Supporting information about constructing student outcome and experience indicators for use in OfS regulation

Review of the selection and grouping of benchmarking factors

Reference OfS 2022.08

Enquiries to ProviderMetrics@officeforstudents.org.uk

Publication date 26 January 2022

Contents

Executive summary **1**
Purpose 1
Our approach to selecting and grouping benchmarking factors 4
Proposed benchmarking factors 6

Methodology used for the review of the selection and grouping of benchmarking factors **10**
Coverage 10
Identifying potential factors and determining the longlist 11
Method for determining the shortlist 12
Method for determining the proposed benchmarking factors 15

Findings from the review of the selection and grouping of benchmarking factors **17**
Findings relevant to identifying potential factors and determining the longlist 17
Findings relevant to determining the shortlist 21
Findings relevant to determining the proposed benchmarking factors 38

Annex A: Statistical modelling methodology **52**

Executive summary

Purpose

1. The Office for Students (OfS) has issued consultations on regulating student outcomes, the Teaching Excellence Framework (TEF) and the construction of student outcome and experience measures for use in OfS regulation.[1] The consultations propose the use of benchmarking to inform the assessments of a higher education provider's performance in respect of the student outcomes and experiences it delivers. The specific purpose and uses of benchmarking are explained in the relevant consultation documents. In each case, benchmarks are used to help interpret a provider's actual performance relative to that of the sector overall, once we have taken into account the mix of students at the provider or the provision being offered, based on a set of benchmarking factors. The consultation documents contain proposals about the factors we have selected to use as benchmarking factors, and how we have grouped them.

2. This document has been published as supporting information alongside the consultations, to aid higher education providers and other stakeholders in understanding the evidence and rationale that has informed our proposals related to benchmarking factors. While we anticipate readers of the consultations will find the information in this supporting document useful for understanding the basis for our proposals, readers should note that this document does not contain proposals for consultation. Our consultation on the construction of student outcome and experience measures for use in OfS regulation invites views on our selection and grouping of benchmarking factors through the consultation questions raised at Proposal 10 of that document (hereafter referred to as the indicators consultation).[2]

What is a benchmark?

3. When calculating student outcome and experience measures as data indicators, each indicator that the OfS calculates represents the outcomes that we have observed for the students at a particular provider at a particular point in time. The calculation of a benchmark gives us a counterfactual for the observed outcomes. This enables our regulation of student outcomes and the TEF to understand a provider's performance in relation to the higher education sector as a whole.[3]

4. In making these comparisons, the proposals we make in proposal 10 of the indicators consultation aim to take account of factors which describe the profile of students and provision delivered by higher education providers and are correlated with the outcomes and experiences we are measuring. The benchmarking methodology we use then makes use of

[1] See www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/.

[2] See www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/outcome-and-experience-data/.

[3] More generally, benchmarks can also be used to assess similarities between individual providers.

the unique combinations formed by the selected student and course characteristics acting as benchmarking factors: we refer to these unique combinations as benchmarking groups.[4]

5. This document describes the evidence and rationale that has informed the selection and grouping of benchmarking factors we have proposed in the indicators consultation, for the student outcome and experience measures for undergraduate students. It has been informed by the guiding principles proposed in Annex D of the indicators consultation.[5] These principles helped us reach balanced views of the policy considerations and statistical properties associated with a range of candidate benchmarking factors. Achieving an appropriate balance in this regard has been critical to ensure that our proposed approach results in a method that remains robust for generating meaningful benchmarks. For this reason, this document combines a narrative about our judgements of each factor's suitability for benchmarking, and evidence used to assess the statistical properties of each candidate factor using results from statistical models. As explained further in paragraphs 7 to 11, inclusion of all possible factors is not possible without compromising the integrity of the method and hence the value of its outputs.

6. Factor selection for the measures for postgraduate students is not considered here. As described in our related consultations, we expect to introduce benchmarking of provision at postgraduate levels of study in future and will share equivalent analysis that informs factor selection for that purpose.

Why is benchmarking factor selection important?

7. Selection of an appropriate range and number of benchmarking factors is critical for the construction and use of meaningful benchmarks, capable of supporting the assessments of provider performance we anticipate making. Benchmarks give information about the values that the sector overall might have achieved for the indicator, if the characteristics included as benchmarking factors are the only ones that are important. In doing so, the 'effect' of each factor included in calculation of the benchmark is, to all intents and purposes, neutralised. Where differences exist between an indicator and its corresponding benchmark, these may be due to the provider's performance, or they may be due to some other characteristic which is not included in the benchmark. In the context of our anticipated uses it is undesirable that differences between an indicator and its corresponding benchmark are driven by characteristics which have a large effect on the outcome we are measuring and are consistent with our policy intention, but have not been selected for inclusion as benchmarking factors.

8. However, it is not possible for benchmarks to take account of every characteristic that is correlated with the outcome we are measuring. We need to ensure, as far as possible, that each benchmarking group refers to a reasonable number of students from multiple providers.

[4] The benchmarking methodology is described further in proposal 10 in the consultation about the construction of student outcome and experience measures, with further statistical detail of its implementation described in the statistical methods supporting document. Both are available at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/outcome-and-experience-data/.

[5] See Annex D the consultation about the construction of student outcome and experience measures at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/outcome-and-experience-data/.

This is because we need to mitigate the risks of 'self-benchmarking', as explained in the indicators consultation and in our supporting document about the statistical methods we use. Self-benchmarking can arise when the characteristics of students at the provider in question do not frequently occur among students in the wider sector, and the provider's own students would be making a substantial contribution to its benchmark. Where this happens the utility of the benchmark is undermined. Selection of benchmarking factors therefore needs to minimise the occasions on which we might encounter this situation, by forming benchmarking groups that each refer to a reasonable number of students from multiple providers.

9. The grouping of the attributes of a selected benchmarking factor is an important part of this process.[6] The basis on which we define the benchmarking groups is key to the integrity and robustness of the benchmarks we calculate. As noted in paragraph 5, the definition of benchmarking groups relies on achieving an appropriate balance of the policy considerations and statistical properties associated with a range of candidate benchmarking factors. The guiding principles we have proposed in Annex D of the indicators consultation recognise the tensions involved in this, and are comprised of principles which focus on policy considerations alongside those focused on statistical properties.[7]

10. Application of the principles means that a range of candidate factors have been tested for their fit with the whole set of proposed principles, with the factors proposed in the indicators consultation representing those which we consider offer the best overall fit. To develop those proposals we used the principles to first identify the credible factors, and then the preferred ones. Throughout this document, the term 'principle' always refers to the set of principles outlined in Annex D of the indicators consultation, and we use the shorthand that follows in order to simplify references to specific principles:

    a. The set of **core principles** (as set out in paragraph 4 of Annex D) relating to the factors considered for selection broadly cover:

        i. The applicability to and availability for all types of providers across England.

        ii. The use of the factor should neutralise the effect of characteristics only where this is consistent with policy objectives.

        iii. The factor should refer to characteristics of students or courses which explain the context of the provider and are outside of its control, or undesirable for it to control for.

        iv. The need to preserve the statistical integrity of the benchmarking approach.

[6] Where we refer to attributes of a factor, we mean the individual categories within the factor. While some student and course characteristics have only a small number of attributes, others have a much larger number of categories, which may be more or less distinct from one another. For example, the factor 'eligibility for free school meals' has two distinct attributes, one labelled 'eligible' and another labelled 'not eligible', in addition to an 'unknown' category. Whereas the factor 'entry qualifications' has potentially hundreds of attributes which might be less distinctive depending on the level of detail at which we consider A-level and other prior attainment grade profiles.

[7] See Annex D the consultation about the construction of student outcome and experience measures at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/oucome-and-experience-data/.

b. Principles that relate to the **availability and data quality** of the factors.

c. Principles that relate to the **statistical properties** of the factors.

d. Principles that relate to defining **groupings of the attributes** within the factor.

11. In the selection of benchmarking factors proposed in the indicators consultation, we have considered the use of consistent approaches to benchmarking factors across our measures wherever possible and appropriate. However, we do not consider it appropriate to do so at the expense of the statistical integrity of the method or its ability to calculate meaningful benchmarks. Rather, we have prioritised those factors which demonstrate the strongest correlations with the outcomes, where these also satisfy our principles for benchmarking. The use of factors which identify material differences in student outcomes and experiences across the higher education sector will help to ensure that the benchmarks we construct are effective in providing meaningful information for the purposes of interpreting individual providers' performance.

Our approach to selecting and grouping benchmarking factors

12. This document and its annexes describe the approach we have taken in the indicators consultation to selecting and grouping the proposed benchmarking factors for indicators reporting on the student outcomes and experiences of undergraduate students, for each mode of study. We summarise the proposed approach taken in that consultation here.

13. Our selection of benchmarking factors started by establishing a comprehensive list of potential factors that could be drawn from existing data sources, which is one of the principles on availability. We tested the fit of each factor with a subset of the principles to remove factors that were clearly not credible candidates where we considered that one or more of the following meant that at least one of the core principles were not met:

    a. The data quality of the factors.

    b. The extent to which the factors are applicable and available for all types of providers across England.

    c. The extent to which the factor is limited to a small group of students or subsection of providers across England.

    d. The extent to which the factors are entirely within a provider's control.

    e. The extent to which the definition of factors overlap with other potential factors.

14. After removing factors that were not credible candidates, the remaining potential factors were included in initial statistical models for each of the proposed student outcomes and experience measures (the definitions and coverage of these measures directly align with the proposals made in our related consultations). Separate statistical models were constructed for each measure and by mode of study. These models seek to identify whether and how factors are correlated with differences in performance for each of the measures and modes of study.

15. We then used a combination of the results from the initial statistical models and a consideration of best fit with a wider range of principles to identify a shortlist of candidate factors. Where necessary, we also used further statistical modelling that looked more closely at the correlations of particular factors. We considered:

    a. Results from the initial statistical models to understand how well the factors fit with the statistical principles, and which factors were most correlated with the modelled outcome.

    b. How far factors were measuring similar properties to one another (for example, there are various measures of students who are underrepresented[8] in higher education or from disadvantaged backgrounds), in line with the core principles for benchmarks.

    c. Whether using the factor to neutralise the effect of characteristics is consistent with policy objectives, in line with the core principles for benchmarks. For example, if there is a difference in the outcomes or experiences of students for those who opt to take a sandwich year, to neutralise this impact through benchmarking may risk reducing incentives for providers to offer options that lead to better student outcomes.

    d. The extent to which factors are within a provider's control, or undesirable for it to control for, in line with the core principles for benchmarks. For example, if differences in progression outcomes are correlated with the awarding body, but the awarding body is chosen by the provider, we may consider this within a provider's control.

    e. Results from bespoke statistical models to understand the correlation of a factor. For example, understanding the impact of changing the reference group in the modelling or understanding whether the factor impacted any particular level of study.

16. In general, a factor would need to have a high correlation with the modelled outcome and have not failed on any other principles to be retained on the shortlist. To establish that a factor did not fail on any other principles requires careful consideration of our policy objectives for benchmarking, including whether they are within a provider's control and the incentives created for providers to change their behaviour in terms of the students they recruit or the provision they offer in ways that we consider undesirable. Factors which were most correlated with the outcomes, and which were consistent with our policy objectives were preferred. This prioritisation was necessary to preserve the statistical integrity of the benchmarking approach, which is reflected in the principles. To do this, we also considered

[8] We use the term 'students from underrepresented groups' throughout this consultation. It includes all groups of potential or current students where the OfS can identify gaps in equality of opportunity in different parts of the student lifecycle. In determining the groups falling within this definition, the OfS has given due regard to students who share particular characteristics that are protected under the Equality Act 2010 as well as students who are otherwise underrepresented or disadvantaged. When referring to underrepresented groups, the OfS considers this to include, among others, students from deprived areas, areas of lower higher education participation, or both; some black, Asian and minority ethnic students; mature students; and disabled students (whether or not they are in receipt of Disabled Students Allowance). There are some student groups with protected characteristics under the Equality Act 2010 where the OfS has been prevented from determining whether they are underrepresented at different points of the student lifecycle, because either: data is collected at a national level but there are gaps in disclosure and absence of comprehensive data (for example in relation to religion or belief, sexual orientation and gender reassignment); or data is not collected at a national level (for example in relation to marriage and civil partnership, and pregnancy and maternity).

how attributes within factors could be grouped, in line with the principles on grouping attributes, in order to keep the number of benchmarking groups to the minimum required for the factor to remain effective.

17. The rationale for each factor that we have not included on the shortlist is described in the main body of this document.

18. Using the shortlisted factors only, we re-created the initial statistical models (and where appropriate the bespoke models) to allow final considerations of which factors should be selected for benchmarking. This involved considering:

    a. Results from the recreated statistical models to understand whether the factors still fit with the statistical principles, and which factors were most correlated with the modelled outcome.

    b. How the selection of a particular factor could impact the statistical integrity of the benchmarking approach. To understand this, we calculated the number of distinct benchmarking groups, the number of students in very small benchmarking groups and the level of self-benchmarking.

Proposed benchmarking factors

19. Proposal 10 of the indicators consultation proposes a set of benchmarking factors for each of the student outcome and experience indicators. Those proposals resulted from application of the principles in Annex D of the indicators consultation, and the approach summarised in paragraphs 12 to 18 above. They are considered to represent the best fit with the range of policy objectives and statistical properties for benchmarking. We repeat the proposed benchmarking factors here for ease of reference, in Tables 1 to 4.

Table 1: Proposed benchmarking factors for continuation indicators

Benchmarking factor	Continuation: full-time	Continuation: part-time	Continuation: apprenticeship
Level of study (First degree, other undergraduate, undergraduate with postgraduate components)	✓ (Other undergraduate separated into that at Level 4 and that at Level 5+)	✓	✓
Subject of study (CAH level 1 groups)	✓	✓	✓
Entry qualifications	✓ (11 groupings)	✓ (5 groupings)	✓ (5 groupings)
Expected course length (Expected course length of less than a year, or otherwise)	✗	✓	✗
ABCS quintile	✓	✓	✓

Benchmarking factor	Continuation: full-time	Continuation: part-time	Continuation: apprenticeship
(Continuation ABCS Quintiles 1 to 5 for the relevant mode of study, non-UK domiciled)			
Total distinct benchmarking groups	5,544	3,780	1,890

Table 2: Proposed benchmarking factors for completion indicators

Benchmarking factor	Completion: full-time	Completion: part-time	Completion: apprenticeship
Level of study (First degree, other undergraduate, undergraduate with postgraduate components)	✓ (Other undergraduate separated into that at Level 4 and that at Level 5+)	✓	✓
Subject of study (CAH level 1 groups)	✓	✓	✓
Entry qualifications	✓ (11 groupings)	✓ (5 groupings)	✓ (5 groupings)
Expected course length	✓ (Expected course length of less than two years, two years, or at least three years)	✓ (Expected course length of less than a year, or otherwise)	✗
ABCS group (Completion ABCS groups, or non-UK domiciled)	✓	✓	✓
Total distinct benchmarking groups	16,632 (2,772 without ABCS quintiles)	3,780 (630 without ABCS quintiles)	1,890 (315 without ABCS quintiles)

Table 3: Proposed benchmarking factors for progression indicators

Benchmarking factor	Progression: full-time	Progression: part-time	Progression: apprenticeship
Year qualification obtained	✓	✓	✓
Level of study (First degree, other undergraduate, undergraduate with postgraduate components)	✓	✓	✓
Subject of study	✓ (CAH level 2 groups)[9]	✓ (Broadly defined subject groups)	✓ (Broadly defined subject groups)
Entry qualifications	✓ (11 groupings)	✓ (5 groupings)	✓ (5 groupings)
ABCS group (Progression ABCS group)	✓	✓	✓
Geography of employment quintile (Quintile 1, Quintiles 2 and 3, Quintiles 4, 5 and unknown)	✓	✓	✓
Total distinct benchmarking groups[10]	80,784 (13,464 without ABCS quintiles)	10,800 (1,800 without ABCS quintiles)	10,800 (1,800 without ABCS quintiles)

[9] For benchmarking purposes, the CAH level 2 group for Celtic studies (CAH19-02) has been combined into the Languages and area studies group (CAH19-04).

[10] The total number of benchmarking groups for progression measures reflects the four years of GO survey responses that will be used in the construction of student outcomes indicators in steady state. This analysis (and the illustrative data released alongside the consultation on constructing student outcome and experience measures) makes use of the two years of GO survey responses that are currently available.

Table 4: Proposed benchmarking factors for student experience indicators

Benchmarking factor	Student experience: full-time	Student experience: part-time	Student experience: apprenticeship
Year of survey	✓	✓	✓
Level of study (First degree, other undergraduate, undergraduate with postgraduate components)	✓	✓	✓
Subject of study	✓ (CAH level 2 groups)[11]	✓ (Broadly defined subject groups)	✓ (Broadly defined subject groups)
Age on entry (Under 21 or unknown, 21 to 30, 31 and over)	✓	✓	✓
Disability (Disability reported, no disability reported)	✓	✓	✓
Ethnicity (Asian, Black, Mixed, Other, Unknown or White, non-UK domiciled)	✓	✓	✓
Sex (Female or other, Male)	✓	✗	✗
Total distinct benchmarking groups	29,376	4,320	4,320

[11] For benchmarking purposes and the definition of split indicators, the CAH level 2 group for Celtic studies (CAH19-02) has been combined into the Languages and area studies group (CAH19-04).

Methodology used for the review of the selection and grouping of benchmarking factors

20. This section of the document describes in more detail the process we followed to identify the proposed benchmarking factors and includes further details on the method.

Coverage

21. Analysis to inform the selection of benchmarking factors considered each of the student outcome and experience indicators in turn. Specifically, those reporting on the rates of continuation in, completion of and progression from higher education study, as well as student experience indicators based on responses to the NSS. The definitions of each measure are aligned to those proposed through our consultations. When considering benchmarking factors, we examined the combination of indicator and mode of study, with the following exceptions:

    a. The potential for conducting the appropriate statistical modelling is more limited for the apprenticeships mode of study due to the more limited spread and characteristics of apprenticeship students across the sector. When considered at the level of detail necessary within the models, there are currently insufficient student numbers for those models to be robust. Therefore, this section does not include any statistical modelling by indicator for apprenticeship students. We have considered how applicable any factors were to apprenticeship students throughout the benchmarking factor selection process and include analysis related to the risks of self-benchmarking by indicator for apprenticeship students.

    b. For the student experience indicators, the analysis has focused on the five scales that are proposed for use in the TEF consultation. Initial analysis of all NSS scales showed similar patterns for the other scales.

    c. For the completion indicators, the analysis focused on the cohort tracking indicator. Because the approach to constructing the compound indicator differs to all other indicators it requires a different more complex formulation for the statistical models. We consider that factors selected on the basis of results from the cohort tracking indicator will be equally applicable to the compound indicator. This is because they are both indicators of completion outcomes, and our aim has been to select factors which are correlated with the outcome itself, and robust to the specifics of how we might choose to measure that outcome. Initial analysis of the applicability of the factors selected for the compound indicator showed similar patterns.

22. For each indicator, we aggregated the four most recent years of available data (or two years for progression, using the currently available GO survey data). This aggregation of years is aligned with the approach proposed in our consultations.

23. In this review, we have included students across the sector who are undergraduate students registered at OfS-registered providers, as of 21 December 2021.[12] In our view, benchmarking providers that are subject to OfS regulation against those that are not, would not be appropriate

[12] The definition of an undergraduate student directly aligns with our proposals.

in the context of our regulatory objectives. We also take the view that benchmarks used in regulation of student outcomes and in the TEF (for English providers) should be consistent.[13]

Identifying potential factors and determining the longlist

24. Our selection of benchmarking factors started by establishing a comprehensive list of potential factors that could be drawn from existing data sources, which is one of the principles on availability. Based on this principle, the potential factors considered in this review were sourced from:

    - Higher Education Statistics Agency (HESA) Student and Student Alternative records
    - The Education and Skills Funding Agency's (ESFA) Individualised Learner Record (ILR) and Learning Aim Reference Service (LARS)
    - The Department for Education's National Pupil Database (NPD)[14]
    - Student Loans Company (SLC) data
    - The National Student Survey (NSS)
    - The Graduate Outcomes (GO) survey
    - The National Statistics Indices of Multiple Deprivations publications
    - Office for Students data on higher education participation or outcomes by area or student characteristics.

25. We tested the fit of each potential factor with a subset of the core principles to remove factors that were clearly not credible candidates to create a longlist of potential factors. Factors were retained as potential factors where none of the core principles were likely to be significantly compromised. We considered:

    a. The data quality of the factor. Factors were tested against the criteria included within the OfS data quality framework.[15]

    b. The extent to which the factor is applicable and available for all types of providers across England. For example, factors were not considered credible candidates where information was not available for further education colleges.

    c. The extent to which the factor is limited to a small group of students or subsection of providers across England. For example, the number of students reported as indicating

[13] See proposal 10 in the consultation about the construction of student outcome and experience measures at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/outcome-and-experience-data/.

[14] Data from the NPD is supplied from the Department for Education (DfE). The DfE does not accept responsibility for any inferences or conclusions derived from the NPD data by third parties.

[15] See www.officeforstudents.org.uk/publications/differences-in-student-outcomes-further-characteristics/.

whether a student has been in care (care experienced) is not considered a credible candidate because 1.1 per cent of full-time UK domiciled undergraduate entrants were reported as such in 2019-20.[16]

d. The extent to which the factor is within a provider's control. For example, the degree classification awarded is not considered a credible candidate for benchmarking completion and progression indicators.

e. The extent to which the definition of the factor overlaps with other potential factors. For example, domicile is not considered as credible given the overlap with home region by NUTS. Significant overlap between factors is not desirable for the final selection of benchmarking factors and risks collinearity in the statistical modelling.

26. In some cases, the fit of factors may differ by indicator or by mode of study. The full list of potential factors and the rationale for any factor not making the longlist are described in our findings.

Method for determining the shortlist

27. Having established the longlist of factors we considered to be potential candidates for benchmarking factors, statistical modelling was used to test the longlist of potential candidate factors against the statistical principles, as outlined in Annex D of the indicators consultation. We also considered the longlisted factors against the policy objectives also outlined in that annex, to establish a shortlist of candidates which represented a reasonable fit with all of the principles.

28. A detailed description of the statistical modelling methodology is provided in Annex A of this document, its purpose being to ensure the factor is correlated with the student outcome in question, once other factors are controlled for. For example, if a factor was an otherwise suitable benchmarking factor, but showed no statistical relationship with the student outcome, then there would be no value in benchmarking by this factor.

29. As part of this modelling, we calculated estimated differences in outcomes between attributes within the factors, after controlling for the other factors. Although these estimated differences do not imply a causal link between the factor and the modelled outcome, their size (and statistical significance) is important to consider. If these differences were not significant, trivial in size, or much smaller than for other factors, then the factor would be of a lower priority for inclusion in benchmarking. Estimated differences that relate to attributes that are 'unknown', 'N/A' or similar were not used to justify the inclusion of a factor, as these indicate that it is the coverage of the factor that is of interest, rather than information provided by the factor itself.

30. The differences were estimated from statistical models with uncertainty, which is indicated by 95 per cent confidence intervals. These indicate how much (observable) uncertainty there is around a given statistic: we would expect the true value of the statistic to lie between the bounds of the intervals 95 per cent of the time, given the data in front of us. In other words, we are 95 per cent 'confident' that the true value lies between the two intervals. The 95 per cent

[16] See www.officeforstudents.org.uk/publications/equality-diversity-and-student-characteristics-data-2010-11-to-2019-20/.

significance level was primarily chosen to be illustrative of the observable statistical uncertainty. It also provides a tolerance of 'false negative' errors that suits our uses on this occasion, based on our expert judgement. A false negative would occur if we incorrectly dismissed an actual difference in performance for an attribute because our modelling suggested it was not statistically significant. We have only reported estimated differences when they are significant at the 95 percent level. This can be interpreted as us being 95 percent 'confident' that the difference is not zero (i.e. that there is a true difference in performance between the attribute in question and the reference attribute, after controlling for the other factors in the model).

31. We have made no adjustment for multiple comparisons when calculating these confidence intervals. This means that we should expect that roughly 5 per cent of the estimated differences in this analysis that appear to be statistically significant (because their confidence intervals do not overlap with zero) exist purely by random chance ('false positives').[17]

32. The estimated differences in outcomes were calculated relative to the reference group within each factor, which is generally chosen to either be the largest group, or an attribute expected to have distinct outcomes from the others. A different choice of reference group may give different results, so this was taken into consideration, with further modelling conducted where further understanding was sought.

33. The estimated difference for each attribute shows the effect of belonging to that group rather than the reference group. The larger these numbers are across the attributes of a given factor, the stronger the statistical relationship between that factor and the outcome. For example, a -1 per cent difference for female students would suggest, all other things being equal, that the effect of being female is to have a 1 per cent greater chance of a successful outcome than if the same students were male.

34. The results of this initial modelling are available separately for each indicator and mode of study, in the accompanying results workbooks.[18]

35. As noted in paragraph 27, alongside the results of the initial statistical modelling of the longlisted factors, the factors have been considered against all of the benchmarking principles. We made judgements to establish each factor's best fit with the wider range of principles. Factors may not have been taken forward at this stage based on a poor fit with one or more of the principles, but common reasons included:

    a. Using the factor to neutralise the effect of characteristics is considered inconsistent with policy objectives.

[17] The OfS considers whether to make an adjustment for multiple comparisons separately in respect of each set of statistics and analysis we produce. For example, our consideration of multiple comparisons in respect of the proposed construction and presentation of student outcomes and experience indicators is described in our statistical methods document available at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/outcome-and-experience-data/.

[18] See sheet 1 of the appropriate results workbook (workbooks are published separately for each indicator and mode of study) at www.officeforstudents.org.uk/publications/review-of-the-selection-and-grouping-of-benchmarking-factors/.

b. The factor is within a provider's control, and not desirable to control for.

c. The statistical modelling suggests that the factor is less correlated with the outcome or was considered of lower priority than other factors that measure similar concepts. For example, when considering the various indicators which aim to identify students from disadvantaged backgrounds, we would consider taking forward those that appear to have the strongest correlation with the modelled outcome. This prioritisation helps to limit the effect of collinearity on future modelling.

d. The modelled differences in outcomes across attributes within a factor are of a trivial size or the factor does not significantly improve the model fit.

36. To aid these judgements, bespoke statistical modelling has been conducted to check that:

a. Our view of the relative priorities of the inclusion of factors does not change as factors are removed from the models. For example, we checked whether the estimated differences between attributes changed for the remaining measures of disadvantage, once we removed some measures of disadvantage as factors in the models. This further modelling gives us confidence that the results we are using are not obscured by collinearity in our initial models.

b. Results are not materially different when modelling is conducted separately for each level of study. For example, it may have been the factors we identified as high priority based on models containing all undergraduates would be different to those identified as high priority based on modelling of just the 'other undergraduate' level of study. As there are benefits of simplicity and ease of understanding in using the same benchmarking factors across all undergraduate levels, however, we would need to see substantial differences to propose the use of different factors for each level. This check also allowed us to investigate differences between levels of study when we incorporated information on the academic level of the qualification (FHEQ) within each level of study – for example, to see if there was benefit in differentiating between FHEQ Level 4 and Level 5 provision in our benchmarking.

c. Factors that do not appear to significantly improve model fit, as main effects also do not have any particularly large or important effects when interacted with any of the other factors.

d. Our view of the relative priorities for the inclusion of factors does not change with a change of reference group, particularly in cases where there was more than one reasonable option to choose between when setting the reference group, as described in paragraph 32.

37. Once we had a sufficient understanding of the relative merits of each of the longlisted candidate factors, we considered different ways of grouping attributes within each factor before determining whether and how to shortlist it. The aim was to maximise the information included in the benchmarks without risking the statistical integrity of the benchmarking process. We were guided in this process by our principles for defining groupings of the attributes within the benchmarking factors (as outlined in Annex D of the indicators consultation) and, in particular, aimed to:

a. Create coherent groupings that make practical sense and share a qualitative similarity.

b. Avoid combining groups with distinct outcomes or behaviours with respect to the indicator in question.

c. Use previously established groupings where possible.

d. Minimise the number of groups with small numbers of students.

e. Aggregate factors as far as possible within these constraints to create the minimum number of categories necessary for the benchmarking factor to be effective.

38. Once we found aggregations that balanced these competing priorities, we assessed whether factors were still worthy of shortlisting. In some cases, the factor may become a lower priority for including in the shortlist. For example, we might consider that a factor becomes a lower priority if its only large, modelled differences related to very small groups that could not be aggregated without creating incoherent or disparate groupings and masking the statistical relationships that had been identified.

Method for determining the proposed benchmarking factors

39. This section describes our approach to identifying the benchmarking factors per indicator and mode of study from the shortlist, as set out in the indicators consultation. This included consideration of results of the statistical models and the consideration of best fit with the wider range of principles.

40. The accompanying results workbooks include the detailed modelling results for the shortlisted factors for each indicator and mode.[19] At this stage, the results from the previous shortlisting section were used to inform groupings of attributes; these groups have been used for modelling the shortlist of factors.

41. In identifying the proposed benchmarking factors, we primarily considered the principle that the number and definition of the factors should not compromise the statistical integrity of the broader benchmarking approach. When evaluating potential sets of factors at this stage, we considered to what extent this principle may be breached. To assess this, we considered:

    a. How many distinct benchmarking groups could theoretically be created?

    b. How many distinct benchmarking groups with at least one student would be created?

    c. How many students would be allocated to small benchmarking groups (i.e. those with five or fewer students, or 20 or fewer students)?

    d. What would the contribution to benchmark be for each provider?

[19] See sheet 2 of the appropriate results workbook (workbooks are published separately for each indicator and mode of study) at www.officeforstudents.org.uk/publications/review-of-the-selection-and-grouping-of-benchmarking-factors/.

42. Due to the different groupings of attributes, and the reduced number of factors included in the shortlist models, results are likely to differ from those observed at the shortlisting stage. In some cases, for example for entry qualifications, small groups with larger estimated differences have been incorporated into larger groups with smaller estimated differences. This may mean that the maximum estimated differences between attributes appear smaller than before, but often they will relate to larger student populations.

Findings from the review of the selection and grouping of benchmarking factors

43. This section of the document describes in more detail our findings at each stage of the methodology described in paragraphs 21 to 42 above.

Findings relevant to identifying potential factors and determining the longlist

44. The potential factors we have identified from the data sources listed in paragraph 24 above are listed in Tables 5 to 7 below, which indicate whether each factor was included on the longlist, and, where factors were not included, the rationale for this. In summarising our judgements at this longlisting stage, we have grouped the candidate factors into three broad groups:

    a. Protected characteristic factors

    b. Other student characteristic factors

    c. Characteristics of study and other structural factors.

45. The potential factors we have identified from the data sources listed in paragraph 24 and grouped as a protected characteristic are listed in Table 5.

Table 5: Potential factors based on protected characteristics

Factor	Data source	On the longlist?	Rationale for not being on the longlist
Age on entry	HESA Student and Student Alternative records, ILR	Yes	N/A
Disability type	HESA Student and Student Alternative records, ILR	Yes	N/A
Ethnicity	HESA Student and Student Alternative records, ILR	Yes	N/A
Gender reassignment[20]	HESA Student and Student Alternative records	No	Insufficient data quality. Currently not available for further education colleges.
Marital status	None	No	Not available
Pregnancy and maternity	None	No	Not available
Religious belief	HESA Student and Student Alternative records	No	Insufficient data quality. Currently not available for further education colleges.

[20] HESA returns collect information relevant to gender reassignment by asking whether, according to their own assessment, a student's gender identity is consistent with their sex registered at birth.

Factor	Data source	On the longlist?	Rationale for not being on the longlist
Sex	HESA Student and Student Alternative records, ILR	Yes	N/A
Sexual orientation	HESA Student and Student Alternative records	No	Insufficient data quality. Currently not available for further education colleges.

46. The potential factors we have identified from the data sources listed in paragraph 24 and grouped as other student characteristics are listed in Table 6.

Table 6: Potential factors based on other student characteristics

Factor	Data source	On the longlist?	Rationale for not being on the longlist
Adult HE quintile, from the 2011 census	OfS data on higher education participation or outcomes by area	Yes	N/A
Associations between characteristics of students (ABCS) quintile	OfS data on higher education participation or outcomes by student characteristics.	Yes	N/A
Care experience	HESA Student record	No	Not sufficient data quality
Carers	None	No	Not available
Children of military families	None	No	Not available
Deprivation quintile (IMD) (and country)	National Statistics Indices of Multiple Deprivations publications	Yes	N/A
Eligibility for free school meals	NPD	Yes	N/A
English as a second language	None	No	Not available
Home region of the student, categorised by NUTS3	HESA Student and Student Alternative records, ILR	Yes	N/A
Household residual income	SLC	Yes (full-time only)	N/A for full-time For part-time, partial data coverage.
Income deprivation affecting children index (IDACI)	National Statistics Indices of Multiple Deprivations publications	Yes	N/A

Factor	Data source	On the longlist?	Rationale for not being on the longlist
Key stage 4 – maintained school status	NPD	Yes (full-time only)	N/A for full-time For part-time, insufficient population on account of more limited potential for identifying an NPD record for the student.
Parental education	HESA Student record	No	Partial data coverage. Currently not available for further education colleges or providers submitting the HESA Student Alternative record.[21]
Participation of local areas (POLAR4)	OfS data on higher education participation or outcomes by area	No	Factor is no longer being updated and is based on entrants between 2009-10 and 2013-14.
Key stage 4 – school region, categorised by NUTS3	NPD	No	Significant overlap with the definition of home region of the student risks collinearity.
Refugees	None	No	Not available
Student estranged from parents	SLC	Yes (full-time only)	N/A for full-time For part-time, insufficient population.
Socioeconomic background	HESA Student record[22]	Yes (full-time only)	N/A for full-time For part-time, partial data coverage.
Tracking underrepresent-ation of areas (TUNDRA) quintile	OfS data on higher education participation or outcomes by area	Yes	N/A

47. The potential factors we have identified from the data sources listed in paragraph 24 and grouped as characteristics of study and other structural factors are listed in Table 7.

[21] Parental education was introduced to the HESA Student Alternative record for the 2020-21 academic year but was not available for the years considered in this analysis.

[22] Note that although the analysis presented here was limited to this information through the HESA Student record, equivalent information is collected for the HESA Student Alternative and ILR records. However, the information available from these sources is not yet deemed to have sufficient data quality. Nevertheless, this wider data coverage is the basis for this factor remaining on the longlist.

Table 7: Potential factors based on characteristics of study and other structural factors

Factor	Data source	On the longlist?	Rationale for not being on the longlist
Awarding body	HESA Student and Student Alternative records, ILR and LARS	Yes	N/A
Awarding level of study	HESA Student and Student Alternative records, ILR	No	Significant overlap with the definition of FHEQ level risks collinearity.
Degree classification	HESA Student and Student Alternative records, ILR	No	Entirely within a provider's control
Entry qualifications	HESA Student and Student Alternative records, ILR, NPD	Yes	N/A
Expected course length in years	HESA Student and Student Alternative records, ILR	Yes	N/A
First degree course with integrated foundation year	HESA Student and Student Alternative records, ILR	Yes (full-time only)	N/A for full-time For part-time, insufficient population and data quality.
Geography of employment quintile	OfS data on higher education participation or outcomes by area	Yes	N/A
Level of study (including FHEQ level)	HESA Student and Student Alternative records, ILR	Yes	N/A
Region of study (NUTS)	HESA Student and Student Alternative records, ILR	Yes	N/A
Sandwich year course	HESA Student and Student Alternative records, ILR	Yes (full-time only)	N/A for full-time For part-time, insufficient population.
Subject of study	HESA Student and Student Alternative records, ILR and LARS	Yes	N/A
Study location type (identifying distance learning and local students) and commuting students	HESA Student and Student Alternative records, ILR	Yes	N/A
Year	HESA Student and Student Alternative records, ILR	Yes	N/A

Findings relevant to determining the shortlist

48. The findings that follow in this section reflect the judgements we have made about which of the longlisted factors to include in our shortlist of potential benchmarking factors. They describe the consideration we have given to each factor, following the method described in paragraphs 27 to 38 in this document.

49. We assessed the modelling based on the longlist and identified factors which had statistically significant estimated differences. This is illustrated in Table 8 below. The section that follows Table 8 then considers the fit of each factor with our proposed principles for benchmarking factors. Throughout the commentary we provide in this section of the report, the descriptions we give are summative, and should not be considered as exhaustive of our considerations of each factor.

Illustrating the statistical modelling of the longlisted factors

50. Table 8 below summarises the maximum size of the estimated differences (for the attribute with the largest significant difference in each case) for each of the factors included on the longlist, by indicator and mode of study. While Table 8 reports only the maximum size of the estimated differences per factor, there are significant nuances in the underlying data that cannot be succinctly summarised here. This includes understanding the proportion of attributes with statistically significant estimated differences and the relative population sizes across attributes. Table 8 aims to provide an indication of the relative correlation with the outcome measured. Full results are available for these longlist models in the accompanying results workbooks.[23]

51. The construction and presentation of Table 8 takes the following approach:

    a. Our analysis showed significant consistency of the factors that were correlated across the NSS scales that are used to construct the student experience indicators. As such, only the results from the ‘teaching on my course’ are included for brevity.

    b. Only estimated differences that are statistically significant at the 95 per cent level are included. Where there are no significant differences, the table shows them with a ‘-‘.

    c. Estimated differences which relate to ’unknown’ attributes are not considered for the purposes of identifying the maximum difference.

    d. The maximum size of the estimated difference is shown where the attribute has a population that exceeds 5,000 students.[24]

[23] See sheet 1 of the appropriate results workbook (workbooks are published separately for each indicator and mode of study) at www.officeforstudents.org.uk/publications/review-of-the-selection-and-grouping-of-benchmarking-factors/.

[24] The counts of students used in this analysis include all students used to calculate each indicator. For all measures except progression (where only two years are currently available), the count of students is based on data across a four-year time series. Therefore, populations for a single year of an indicator are likely to be much smaller than those indicated in Table 8.

e. If the maximum size of the estimated difference corresponds to an attribute with a population smaller than 5,000 students, we instead show the largest estimated difference that satisfies the 5,000 students requirement. In addition to reporting this figure, Table 8 also includes the maximum estimated difference for the smaller population in brackets.

Table 8: Maximum estimated differences for the longlisted factors, by indicator and mode of study

Key: Grey shading identifies factors not on the longlist. Peach shading identifies factors on the shortlist. Unshaded identifies a longlisted factor that was not shortlisted.

Maximum estimated differences (percentage points, significant at the 95 per cent level only)	Continuation		Completion (cohort tracking)		Progression		Student experience (teaching on my course scale)	
Factor	Full-time	Part-time	Full-time	Part-time	Full-time	Part-time	Full-time	Part-time
Characteristics of study and other structural factors								
Subject of study (level 2 of the common aggregation hierarchy)	4.4	13.1 (22.0)	6.4	20.2 (24.7)	33.6 (38.0)	15.3 (22.4)	8.1	5.1 (9.4)
Entry qualifications	15.1	23.4	22.2	29.2	11.2 (11.4)	2.4 (81.4)	2.0 (2.2)	-
Level of study (results for FHEQ level)	4.8	2.6 (6.0)	7.8	7.0 (8.5)	3.8 (5.0)	5.5	2.7	-
Year	0.3	3.3	5.2	1.4	2.5	-	4.3	1.2
First degree course with integrated foundation year	4.8		9.4		4.8		1.5	
Expected course length in years	1.7 (9.9)	8.0 (29.0)	7.4 (39.7)	11.8 (68.8)	6.9	- (12.8)	4.4	- (4.8)
Commuting student	3.2	1.8	2.3		1.6	-	-	-
Study location type	1.1 (12.0)	4.8 (12.6)	1.4 (12.7)	7.1 (16.7)	0.6 (11.2)	-	-	-
Geography of employment quintile					13.1	3.2 (5.0)		
Region of study (NUTS)	2.3 (11.6)	14.0	3.6 (10.8)	14.5	6.7	- (18.8)	3.8 (10.6)	-
Awarding body	1.7	4.5	4.2	3.0	-	-	-	- (5.4)
Sandwich year course	-		-		-		-	

Maximum estimated differences (percentage points, significant at the 95 per cent level only)	Continuation		Completion (cohort tracking)		Progression		Student experience (teaching on my course scale)	
Protected characteristics factors								
Age on entry	2.3 (8.4)	2.4 (4.6)	4.7 (12.1)	8.8	3.6 (18.3)	- (10.4)	4.5 (11.8)	- (5.0)
Disability type	2.8	5.5	8.2	5.2 (5.8)	3.6 (9.0)	- (7.6)	2.6	2.6
Ethnicity	0.9	3.0	1.8	3.2	3.1	- (5.9)	3.8	- (1.6)
Sex	1.3	2.6	3.1	2.5	1.8	3.3	1.1	-
Other student characteristic factors								
Associations between characteristics of students (ABCS) quintile	9.0	10.4						
Deprivation quintile (IMD) (and country)	5.7 (9.8)	2.3	3.8 (11.7)	2.3 (14.1)	2.4	- (21.4)	0.7	-
Eligibility for free school meals	1.8	8.4	2.5	- (9.1)	-	-	1.1	-
Adult HE quintile	-	-	-	-	0.9	-	0.3	-
Home region of the student (NUTS)	12.4	2.3	14.4	1.9	3.0	-	1.5	-
Student estranged from parents	2.4		5.9		- (2.9)		2.6	
Household residual income	0.6		1.2		1.4		1.4	
Income deprivation affecting children index (IDACI)	0.6	1.7	1.8	3.1	-	1.6	-	-
Key stage 4 – maintained school status	0.7		-		1.5		1.6	
Socioeconomic background	1.4		2.8 (3.0)		3.5		0.8	
Tracking underrepresentation of areas (TUNDRA) quintile	0.4	-	-	-	-	-	-	-

52. In making the following judgements about which of the longlisted factors have been included in our shortlist of benchmarking factors, we carefully considered each factor following the method described in paragraphs 27 to 38. We provide a commentary in paragraphs 56 to 107 below which summarises our judgement of each of the longlisted factors to construct a shortlist. This commentary is summative and should not be considered as exhaustive of our considerations of each factor nor determinative of the proposals we have made through our consultations.

53. In summarising our judgements at this shortlisting stage, we have grouped the candidate factors into three broad groups:

    a. Characteristics of study and other structural factors

    b. Protected characteristic factors

    c. Other student characteristic factors.

54. Factors within each of these groups are listed in approximate priority order in the discussion that follows, where this ordering is broadly based on the longlist modelling results, bespoke modelling, and our wider judgement of their fit across the benchmarking principles.

55. While each factor was considered on its own merits, it was important to consider the relationships between factors and prioritise in cases where too many factors are measuring similar concepts. To do otherwise would risk collinearity within the shortlist modelling. These considerations are highlighted below, where applicable.

Characteristics of study and other structural factors

Subject of study

56. **Shortlisted (all indicators)**. Across the indicators this factor had many of the largest statistically significant estimated differences, as illustrated in Table 8 above. While the number and size of the differences varied by indicator and mode of study, there was consistently enough evidence to warrant further consideration of the factor, so this factor was shortlisted for all indicators. The inclusion of this factor would allow benchmarking of all student outcome and experience measures to take account of one of the key structural differences in the design and delivery of higher education courses, which affects all higher education providers. As such, shortlisting this factor was considered appropriate and we were confident that, if taken forward as a proposal, its use would be coherent in terms of the policy objectives and incentives related to differences of student outcomes and experiences.

57. To understand how this factor might be implemented if it were taken forward, we considered at this shortlisting stage whether broader or more granular groupings of subject of study were desirable. This factor was initially modelled using level 2 of the Common Aggregation Hierarchy (CAH2), as illustrated through the estimated differences reported in Table 8. We considered whether it was possible to aggregate this factor into broader groups to help support the onward development of an approach which maximised the number of factors (and hence information) accounted for through the benchmarks, without risking the statistical integrity of the benchmarking process or the effectiveness of accounting for the differences the initial modelling indicated across subject areas. When considering the best fit to the principles for grouping attributes within a factor, we concluded that we should use established groupings and:

a. For continuation and completion indicators to use CAH1 groupings. We identified good consistency of observed rates and statistically significant estimated differences across CAH2 within the higher level of aggregation, level 1 (CAH1) in the modelling for continuation and completion indicators.

b. For full-time progression and student experience indicators to use CAH2 groupings. This is because we found that there was not good consistency of observed rates or statistically significant estimated differences within the broader levels of aggregation.

c. For part-time progression and student experience indicators to use the broadly defined subject groups that the OfS has already developed based on further aggregating the CAH1 groupings.[25] This is because the relative populations of the more granular aggregations would risk the statistical integrity of the benchmarking process.

Entry qualifications

58. **Shortlisted (all indicators)**. Across the indicators, Table 8 shows that this factor had some of the largest statistically significant estimated differences. While the number and size of the differences varied by indicator and mode of study, there was consistently enough evidence to warrant further consideration of the factor, so this factor was shortlisted for all indicators.

59. The inclusion of this factor would allow benchmarking to take some account of students who are less well prepared for higher education. It was considered that the importance of doing so was more marked for the continuation and completion measures reporting on entrant cohorts. This is because our regulation of quality and standards requires that an OfS-registered provider must provide all students with the support that they need to succeed in and benefit from higher education.[26] As such, by the time that they feature in indicators reporting on them at the end of their courses, students should have been provided with the appropriate support they required in order to succeed on that course as a whole, regardless of their level of preparation for starting higher education. Differences in student outcomes and experiences at the end of a course may indicate that students did not receive appropriate support and hence may be a proxy for provider performance. The likelihood of this was considered higher in respect of the student experience measures, where it is less clear that the experiences final year students report through the NSS would differ according to students' preparedness to start higher education, other than to reflect upon issues of provider performance.

60. Consequently, we considered that shortlisting this factor was appropriate for continuation and completion measures. We also considered that by retaining this factor within the shortlist for student experience and progression at this stage we have offered benefit of the doubt that the factor may be useful for the purposes of benchmarking student experience and progression indicators, on account of the estimated differences identified through the initial modelling. It

[25] This grouping is included in Annex E of the consultation about the construction of student outcome and experience measures at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/outcome-and-experience-data/.

[26] This requirement is currently formulated as condition B2 of the regulatory framework for higher education in England, available at: www.officeforstudents.org.uk/publications/securing-student-success-regulatory-framework-for-higher-education-in-england/. We are currently consulting on proposed revisions to condition B2. Details of these proposals can be found at www.officeforstudents.org.uk/advice-and-guidance/quality-and-standards/changes-to-our-approach/.

was recognised that use of the factor in benchmarking all measures may guard against incentivising changes in provider's recruitment behaviours that could undermine our ability to promote equality of access and opportunity in English higher education.

61. To understand how this factor might be implemented if it were taken forward, we considered at this shortlisting stage whether broader groupings of entry qualifications were desirable. This factor was initially modelled using a granular derivation of entry qualifications with around 50 attributes. There are no well-established groupings for entry qualifications; retaining this granularity would not be feasible because it would significantly impact the statistical integrity of the benchmarking approach when used in combination with other factors. We considered whether it was possible to aggregate this factor into broader groups to help support the development of an approach which maximised the number of factors (and hence information) accounted for through the benchmarks, without risking the statistical integrity of the benchmarking process or the effectiveness of accounting for the differences the initial modelling indicated across entry qualifications. In considering the best fit to the principles for grouping attributes, we have developed a grouping with 11 categories for full-time indicators and aggregated this further into five categories for part-time and apprenticeship indicators. [27]

Level of study (including FHEQ level)

62. **Shortlisted (all indicators)**. There was consistently enough evidence to warrant further consideration of the factor, so this factor was shortlisted for all indicators. We were confident that, if taken forward as a proposal, its use would be consistent with our policy objectives to measure and report on student outcomes separately for each level of study.

63. Within the initial modelling, models were constructed to cover all undergraduate students. The academic level of the qualification aim according to the FHEQ being included as a factor has meant that level of study was not then explicitly included as a factor in these 'all undergraduates' models, because of its correlation with the FHEQ level and the resulting risks of collinearity. Across the full-time continuation and completion indicators there were statistically significant estimated differences for the FHEQ levels, which led us to shortlist this factor (in addition to level of study) for these indicators. The estimated differences summarised in Table 8 were modelled for all undergraduates using a reference group of FHEQ Level 6, which has a significant proportion of students in comparison to other levels. We considered that there were no policy objectives that would suggest we should not shortlist this factor.

64. However, to better understand the initial modelling results, we ran bespoke modelling to look separately at the different levels of undergraduate qualifications (first degree, undergraduate with postgraduate components and other undergraduate). In particular, bespoke modelling for students at other undergraduate level, and for students on undergraduate courses with postgraduate components, helped us determine that for full-time continuation and completion indicators, there were statistically significant estimated differences for students on other undergraduate level courses according to the FHEQ level of their qualification. This meant that it would be possible to combine the FHEQ level with the other undergraduate level. For these indicators that would create a four-way grouping of other undergraduate at FHEQ Level 4,

[27] These groupings are included in Annex E of the consultation about the construction of student outcome and experience measures at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/outcome-and-experience-data/.

other undergraduate at FHEQ Level 5 or higher, first degree, and undergraduate with postgraduate components. We considered whether it was possible to take a similar approach for other indicators and modes of study, in order to support the development of an approach which maximised the information accounted for through the benchmarks. We established that, in these cases, doing the same thing would risk statistical integrity of the benchmarking process given the relative population sizes of these groups.

Year

65. **Shortlisted (progression and student experience)**. Across the indicators, there were statistically significant estimated differences as illustrated in Table 8. For the progression and student experience indicators the modelling identifies an observable effect which could relate to the impact of the pandemic. Our view was that we should shortlist year for the purposes of benchmarking because, if taken forward as a proposal, this would allow us to account for the potential impact of the pandemic as well as structural and operational changes to the survey instruments that underpin them. It could also offer an appropriate and proportionate approach to taking some account of external influences on progression outcomes which may be related to short term changes in the labour market. We were therefore confident that, if taken forward as a proposal, its use would be coherent in terms of the policy objectives related to differences of student outcomes and experiences.

66. When considering the statistically significant estimated differences for continuation and completion indicators, these were generally fairly small or became small once other factors were removed from the models. Additionally, the current data does not provide evidence of an impact of the coronavirus pandemic in respect of these outcomes. For these reasons, and in the absence of a strong rationale for why year of entry should have a differential impact on these student outcomes across different providers, this factor was not shortlisted for continuation and completion indicators.

First degree course with integrated foundation year

67. **Shortlisted (continuation and completion, full-time)**. Table 8 illustrates that, across the full-time continuation and completion indicators there were statistically significant estimated differences. The size of estimated differences in the initial modelling suggested that there was enough evidence to warrant further consideration of the factor for these indicators. While the inclusion of this factor would allow benchmarking to take some account of students who are less well prepared for higher education, it was considered that the same was true of the entry qualifications factor already shortlisted. We also consider that the inclusion of a foundation year within the structure of a provider's courses is a choice that falls within its control: prospective students have alternative options, such as access to HE courses or further study at Level 3, through which providers may be able to support them to become better prepared for higher education. Accordingly, the factor's fit with the wider benchmarking principles is weakened in these regards. By retaining this factor within the shortlist at this stage we have offered benefit of the doubt that the factor may nonetheless be useful for the purposes of benchmarking full-time continuation and completion outcomes, on account of the estimated differences identified through the initial modelling. We took the view that, if taken forward as a proposal, further modelling of the shortlisted factors would need to examine the correlations between the foundation year and entry qualifications factors. When interpreting that modelling, we would need to give careful consideration to whether accounting for foundation year courses

would be coherent with the policy objectives related to differences of student outcomes and experiences, given the concerns described here.

68. To understand how this factor might be considered in onward development (and, potentially, implementation) of the benchmarking approach for full-time continuation and completion indicators, we considered at this shortlisting stage whether broader or more granular groupings were desirable. We determined that there is no reasonable way to aggregate or disaggregate attributes within this factor into other groupings.

69. While Table 8 shows some statistically significant estimated differences for **progression and student experience indicators**, this factor was not shortlisted for either indicator. We recognise that this factor could help us take some account of students who are less well prepared for higher education, but we do not consider this relevant to indicators that are based on qualifying students. This is because our regulation of quality and standards requires that an OfS-registered provider must provide all students with the support that they need to succeed in and benefit from higher education.[28] As such, students recruited onto first degree courses with integrated foundation years should have been provided with the appropriate support they required in order to succeed on that course as a whole, regardless of their level of preparation for starting higher education. If there are differences in students' progression outcomes and the student experience they report at the **end** of a first degree course that **began** with a foundation year, we consider that such differences may indicate that students did not receive appropriate support and hence may be a proxy for provider performance. Consequently, we consider that shortlisting this factor for progression and student experience indicators would be incoherent with policy objectives related to differences of student outcomes and experiences.

Expected course length in years

70. **Shortlisted (continuation, part-time, and completion)**. Across the indicators there were statistically significant estimated differences as illustrated in Table 8. For the part-time continuation, and the completion indicators, the modelling provides enough evidence to warrant further consideration of the factor for these indicators. The inclusion of this factor would allow benchmarking of completion outcomes to take some account of students who experience life events that may cause them to cease studying. For part-time students who are recorded with an expected course length of less than one year, it would allow benchmarking of continuation outcomes (with a proposed census date of two years after commencement of a higher education course) to take some account of the different motivations and circumstances of these students on shorter courses. As such, shortlisting this factor was considered appropriate. While the expected course lengths of its qualifications is, to some extent, within a provider's control, we have offered benefit of the doubt that the factor may nonetheless be useful for the purposes of benchmarking these indicators, on account of the estimated differences identified through the initial modelling. We took the view that, if taken forward as a proposal, its use would not prove incoherent in terms of the policy objectives related to differences of student outcomes.

[28] This requirement is currently formulated as condition B2 of the regulatory framework for higher education in England, available at: www.officeforstudents.org.uk/publications/securing-student-success-regulatory-framework-for-higher-education-in-england/. We are currently consulting on proposed revisions to condition B2. Details of these proposals can be found at www.officeforstudents.org.uk/advice-and-guidance/quality-and-standards/changes-to-our-approach/.

71. To understand how this factor might be implemented if it were taken forward, we considered at this shortlisting stage whether broader groupings of expected course lengths were desirable. We considered whether it was possible to aggregate this factor into broader groups to help support the development of an approach which maximised the information accounted for through the benchmarks, without risking the statistical integrity of the benchmarking process or the effectiveness of accounting for the differences the initial modelling indicated across course lengths. In considering the best fit to the principles for grouping attributes, we have developed a grouping with three categories for full-time completion indicators and a binary grouping for part-time continuation and completion indicators.

72. While Table 8 shows a statistically significant estimated difference for **full-time continuation**, this factor was not shortlisted for this indicator. This is because estimated differences between attributes were small for attributes with substantial student populations. It was also not considered a good fit with the principle that factors should take account of the effect of characteristics where this is consistent with policy objectives. This indicator looks at the rates of students who continue or qualify one year after starting their course and it is not clear why the expected course length beyond that should impact the likelihood of students continuing beyond the start of that course for reasons that a provider is unable to influence. Any differences may be a proxy for provider performance.

73. While Table 8 also shows some statistically significant estimated differences for **progression and student experience indicators**, this factor was not shortlisted for either indicator. These indicators are derived based on students at the end of their courses and beyond the level of study that we have already incorporated as a shortlisted factor, we consider that this factor was sufficiently within a provider's control, and hence not desirable to control for. If there were differences in the outcomes or experiences of students on qualification based on their course length (above and beyond their level of study), taking account of this through benchmarking may risk reducing incentives for providers to offer options that lead to better student outcomes. For example, it would have the effect of neutralising any differences associated with students studying at the same level of study but with differing expected course lengths based on the inclusion of a sandwich year or year abroad option. Alternatively, differences in outcomes that follow from students experiencing different course lengths may otherwise represent be a proxy for provider performance, which it is not desirable to control for.

Study location type (identifying distance learning and local students) and commuting student

74. **Shortlisted (continuation, full-time)**. There is significant overlap between the definitions used for the derivation of the commuting student and study location type factors. For both factors, across continuation, completion and progression indicators (full-time only), Table 8 illustrates that there were statistically significant estimated differences. The significant attributes were for students that were local, commuting, and distance learning.

75. These factors were not shortlisted for **part-time indicators**, despite some statistically significant estimated differences. This is because the differences were either small and for groups with significantly fewer students than the groups that included distance learner students (which were the reference groups for the models), or they related to groups with very small numbers of students. They were also not shortlisted for the **full-time completion, progression or student experience indicators**, because estimated differences between attributes were small for attributes with substantial student populations. Through bespoke

modelling these differences remained small. Shortlisting this factor was considered appropriate for the **full-time continuation indicator**, where estimated differences were larger, and we were content that, if taken forward as a proposal, its use would not prove incoherent in terms of the policy objectives and incentives related to differences of student outcomes and experiences.

76. When considering the potential implementation of this factor for benchmarking the full-time continuation indicator, we recognised that some of the granular groups contain relatively small numbers of students. There were no established groupings for this factor. When considering the best fit to the other principles of grouping attributes and how to maximise the information included in the benchmarks without risking the statistical integrity of the benchmarking process, we combined these two factors into one and aggregated into broader groups. This means that for the purposes of the full-time continuation indicator, we have formulated a binary grouping of commuting or distance learning, and not commuting or distance learning.

Geography of employment quintile

77. **Shortlisted (progression)**. For the progression indicators, Table 8 illustrates that there were statistically significant estimated differences. We considered that there was enough evidence to warrant further consideration of the factor, so this factor was shortlisted for this indicator. The geography of employment quintiles classify areas across the UK based on graduates who responded to the Graduate Outcomes survey, providing a way to account for the differences between relatively small geographic areas.[29]

78. Taking account of individual geographic areas is not considered feasible for the purposes of defining benchmarking factors. To avoid significant risks of creating large numbers of sparsely populated benchmarking groups leading to self-benchmarking, we would need to consider large areas (such as Government Office regions) which would hide substantial variation. For instance, areas on the East Anglian coast would be counted in the same region as areas just outside London, masking significant differences in these labour markets. We take the view that geography of employment quintiles, classifying areas across the UK based on graduates who responded to the Graduate Outcomes (GO) survey, provides a way to account for the differences between relatively small, and therefore more homogeneous, geographic areas without compromising the integrity of the benchmarking approach.

79. The inclusion of the geography of employment quintile factor would allow benchmarking of progression measures take some account of the likelihood of a graduate's progression outcomes being influenced by the area where the graduates are living and working. We were content that, if taken forward as a proposal, its use would not prove incoherent in terms of the policy objectives and incentives related to differences of student outcomes and experiences.

80. To understand how this factor might be implemented if it were taken forward, we considered at this shortlisting stage whether broader groupings of the quintiles were desirable. In considering the best fit to the principles for grouping attributes, we have aggregated the attributes for the factor into a three-way group of quintile 1, quintiles 2 or 3, and quintiles 4 or 5.

[29] See our November 2021 publication at www.officeforstudents.org.uk/publications/a-geography-of-employment-and-earnings/.

Region of study (NUTS)

81. **Not shortlisted**. Across the indicators, there were statistically significant estimated differences. However, we consider the definition of this factor to be too broad and recognise the risks of collinearity with respect to other area-based factors (including the home region of the student). It is also unclear if the modelling results are indirectly measuring some form of disadvantage, when other shortlisted factors are more direct measures.

82. Moreover, we note that there is a great deal of separation with this factor, i.e., for most providers, most students study within a single region. The modelling controlled for teaching provider, which means that any modelled differences must have been driven by the outcomes of a small minority of students who study away from the provider's main campus. For this reason, we consider that it would not be consistent with policy objectives to, in effect, limit the 'sector' defined for benchmark purposes to only those providers within the same region.

Awarding body

83. **Not shortlisted**. Across the indicators, there were some statistically significant estimated differences that were small or moderate in size. This factor differentiates awards given by the provider, those given under validation arrangements, and those given by Pearson or the Scottish Qualifications Authority. We considered that this factor was too much within a provider's control and its use would not be coherent in terms of the policy objectives related to differences of student outcomes and experiences.

Sandwich year course

84. **Not shortlisted**. Across the indicators, there were either no statistically significant estimated differences or where there were, the estimated differences were small. This factor was not shortlisted because its use would not be coherent in terms of the policy objectives related to differences of student outcomes and experiences. We took this view because we recognise that students who take a sandwich year can find this an enriching experience; taking account of this through benchmarking may risk reducing incentives for providers to offer options that lead to better student outcomes and experience. For example, it would have the effect of neutralising any differences associated with students studying on courses with a sandwich year option. We anticipate the results observed in the modelling may be impacted by the derivation of the factor which relies on the student being on a sandwich year course in the year the indicator is calculated (i.e., on a sandwich year in the year of qualification for the progression indicator) which may not align with the structure of many courses which involve a sandwich year.

Protected characteristic factors

85. The shortlisting of protected characteristic factors is consistent with policy objectives, including our aim to promote equality of opportunity for mature students, disabled students, and black, Asian and minority ethnic students. All of these protected characteristics had statistically significant estimated differences across all indicators, with the exception of sex for part-time student experience indicators. All were considered a good fit across the benchmarking principles and inclusion of the factor would allow benchmarking to take some account of students who are underrepresented in higher education or otherwise prioritised in respect of their protections under the Equality Act 2010. Consequently, all of the following factors were **shortlisted for all indicators**, with the groupings detailed below. To understand how these

factors might be implemented if it were taken forward, we considered at this shortlisting stage whether broader or more granular groupings were desirable. The use of broader groupings may help support the development of an approach which maximised the information accounted for through the benchmarks, without risking the statistical integrity of the benchmarking process.

Age on entry

86. **Shortlisted (all indicators)**. This factor was modelled with granular attributes, one for each age (i.e. age 18 and age 19 separately). There are existing groupings that separate students who are young on entry (under 21) and those who are mature (21 and over). Considering the modelling results across indicators and the coherence with existing groupings we have formulated a three-way grouping of under 21 on entry, 21 to 30 and 31 or older. This grouping still allows the identification of mature students (21 and over).

Disability type

87. **Shortlisted (all indicators)**. This factor was modelled with granular attributes, one for each reported disability type. There are existing groupings that separate students into those who have a disability reported and those who do not. Considering the modelling results across indicators, the relative populations of the individual disability types and the coherence with existing groupings, we have aligned with the existing binary grouping of disability reported and no disability reported. For full-time continuation and completion indicators the modelling results supported keeping the mental health conditions attribute as a separate group. However, given the relatively small student populations associated with this attribute and that we cannot determine whether students with multiple conditions or impairments have a mental health condition, we judged it better to aggregate to the binary grouping described.

Ethnicity

88. **Shortlisted (all indicators)**. This factor was modelled with granular attributes. There are existing groupings that separate students into those who are Black, Asian, Mixed or Other and those who are White. Considering the modelling results across indicators, and the coherence with existing groupings we have aligned with the existing binary grouping of Black, Asian, Mixed or Other and White for the continuation, completion and progression indicators. For the student experience indicators, the modelling results did not support aggregating the attributes in the same way and we note that published NSS results are benchmarking by ethnicity at a more granular level. In shortlisting ethnicity for these indicators, we have not aggregated the attributes into groups.

Sex

89. **Shortlisted (all indicators)**. There is no way to aggregate attributes within this factor into fewer groups.

Other student characteristic factors

90. The remaining longlisted factors all seek to identify student groups that are underrepresented in higher education or students from disadvantaged backgrounds. These factors are:

    a. Adult HE quintile

b. Associations between characteristics of students (ABCS) quintile

c. Deprivation quintile (IMD) (and country)

d. Eligibility for free school meals

e. Home region of the student (NUTS)

f. Household residual income

g. Income deprivation affecting children index (IDACI)

h. Key stage 4 – maintained school status

i. Socioeconomic background

j. Student estranged from parents

k. Tracking underrepresentation of areas (TUNDRA) quintile

91. As described in paragraph 55, when considering how to shortlist these factors it was important to consider how to prioritise amongst them. While the statistical modelling shows statistically significant estimated differences across several of the factors listed in paragraph 90, it was importa90nt to consider the relationships between them and prioritise in cases where too many were measuring similar concepts. It was also important to consider the relationships, and relative priorities of these factors and those reflecting protected characteristics: we have taken the view that shortlisting protected characteristic factors was generally higher priority than these.

92. When considering the size of the statistically significant estimated differences for these factors across the different indicators, we identified that the factors ABCS quintile, IMD and eligibility for free school meals generally had larger estimated differences. We considered the wider fit of these factors to our policy objectives related to differences of student outcomes and experiences and identified these factors as the highest priority for shortlisting. The consideration of the shortlisting potential for each of these higher priority factors by indicator is described below.

Associations between characteristics of students (ABCS) quintile

93. **Shortlisted (continuation only).** The ABCS quintiles are currently only available in relation to the continuation indicator. Across both modes of study, this factor had one of the largest statistically significant estimated differences, meaning that there was enough evidence to warrant further consideration of the factor.

94. ABCS is a set of analyses that seeks to better understand how higher education outcomes vary for groups of students holding different sets of characteristics. We define ABCS quintiles, which represent groups of students on the basis of a set of characteristics, so that we can determine the effect of not just one characteristic on an outcome, but the effect of multiple characteristics.[30] As an intersectional measure, which is designed to differentiate

[30] See www.officeforstudents.org.uk/data-and-analysis/associations-between-characteristics-of-students/.

those individuals with combinations of student and background characteristics that identify them as being least likely to achieve the higher education outcome in question, we consider it is a valuable and effective means of accounting for the material differences that our benchmarking method seeks to account for. The inclusion of this factor therefore be consistent with policy objectives related to differences of student outcomes. As such, we considered it was of high priority to shortlist for continuation indicators.[31]

95. We are not yet able to include ABCS completion or progression quintiles within this analysis in respect of completion or progression outcomes, as the ABCS method has not so far been applied to these outcomes.[32] The OfS expects to extend the ABCS method to produce and publish quintiles for these points of the student lifecycle during 2022. Similarly, we cannot yet include ABCS student experience quintiles within this analysis. In this case, we intend to keep the longer-term benchmarking of student experience indicators under review as our approach to ABCS develops and the outcomes of the current review of the NSS are implemented.[33]

96. To understand how these factors might be implemented if it were taken forward, we considered at this shortlisting stage whether broader or more granular groupings were desirable. The use of broader groupings may help support the development of an approach which maximised the information accounted for through the benchmarks, without risking the statistical integrity of the benchmarking process. This factor was modelled with each quintile separately. There is a reasonable spread of students across each of the quintiles, and in considering the best fit to the principles for grouping attributes, we have chosen not to group attributes within this factor.

Deprivation quintile (IMD) (and country)

97. **Shortlisted (continuation, completion and progression)**. Across the continuation, completion and progression indicators there were statistically significant estimated differences, meaning that there was enough evidence to warrant further consideration of the factor. The shortlisting of this factor is consistent with policy objectives, including our aim to promote equality of opportunity for students from lower household income or socioeconomic status backgrounds. We considered that this area-based measure of these backgrounds was of high priority to shortlist for these indicators.

98. To understand how this factor might be implemented if it were taken forward, we considered at this shortlisting stage whether broader groupings were desirable. The use of broader groupings may help support the development of an approach which maximised the information accounted for through the benchmarks, without risking the statistical integrity of the benchmarking process. This factor was modelled with each quintile separately per country. There are existing binary groupings that combine English IMD quintiles 1 and 2, and

[31] In respect of continuation outcomes, we are able to operationalise use of ABCS quintiles as benchmarking factors immediately, using the analysis published on this methodology each year since 2019.

[32] This means that ABCS completion and progression quintiles have not been included in the construction of the supporting and illustrative data released alongside the consultations.

[33] See www.officeforstudents.org.uk/advice-and-guidance/student-information-and-data/national-student-survey-nss/review-of-the-nss/.

English IMD quintiles 3 to 5. In considering the best fit to the principles for grouping attributes we have aligned with the existing binary grouping of quintiles 1 and 2, and quintiles 3 to 5, for the English deprivation quintiles only.

99. While Table 8 illustrates some statistically significant estimated differences for student experience indicators, this factor was not shortlisted. This is because the estimated differences were small, and we consider that other factors (including protected characteristic factors) were higher priority for shortlisting.

Eligibility for free school meals

100. **Shortlisted (continuation and completion)**. Across the continuation and completion indicators, Table 8 illustrates that there were statistically significant estimated differences. The shortlisting of this factor is consistent with policy objectives, including our aim to promote equality of opportunity for students from lower household income or socioeconomic status backgrounds.

101. To define this factor, we rely on linking HESA and ILR student records to NPD School Census data about pupils at key stage 4. NPD data is available in respect of 2009-10 onwards and has information on pupils attending maintained schools in England. This definition means that there are larger numbers of students classified with an unknown eligibility for free school meals when indicators are based on more historic data, particularly for the completion indicators and for part-time students. For instance, part-time students tend to be more mature which means making a successful link to the available NPD data is less likely because they are further away from school age and may have changed their postcodes or names (for example, through marriage). For the completion indicators, the earliest year of entrant data informing the measure is 2010-11, which means that many students can't be linked to the NPD data. The full-time continuation indicator is not limited in the same way.

102. Notwithstanding the limitations described in paragraph 101, we considered that the coverage is sufficient to support this individual-level measure of students from lower household income or socioeconomic status backgrounds as a high priority to shortlist for the continuation and completion indicators, given its coherence with the policy objectives. We took the view that, if taken forward as a proposal, the groupings used would need to take appropriate account of the potential partiality of its coverage. This would mean including students with unknown eligibility for free school meals as separate group.

103. While some statistically significant estimated differences are illustrated in Table 8 for student experience indicators, this factor was not shortlisted for either progression or student experience indicators. This is because the estimated differences were small, and through bespoke modelling remained small.

Assessment of the remaining factors

104. The following factors were **not shortlisted**, on the basis of concerns about the applicability of the factors for identifying students who are underrepresented in higher education or from disadvantaged backgrounds:

    a. **Adult HE quintile**: The applicability concern here related to the currency of the measure, which is derived from the 2011 census. Moreover, there were either no statistically

significant estimated differences as illustrated in Table 8, or where there were, the estimated differences were very small.

b. **Home region of the student (NUTS)**: The applicability concern here related to the broad definition of the factor and the geographical areas it refers to and we have not identified a viable solution to reducing the breadth of the attributes defined by this factor. Across the indicators, there were statistically significant estimated differences. However, it is unclear if these differences represent this factor acting as an indirect measure (or proxy) of disadvantage. We consider that this lack of clarity, and the availability of more clearly defined and direct measures of underrepresentation and disadvantage within the factors already shortlisted, supports our view that this factor is not shortlisted.

105. While there were some statistically significant estimated differences for the factor **students estranged from parents** across various indicators, this factor identifies relatively small groups of students. We consider that these group sizes could risk the statistical integrity of the benchmarking process once combined with other factors. The factor was also considered to overlap with other, higher priority, student characteristics already included in the shortlist. Taken together, we considered that this meant that inclusion of the factor would risk the statistical integrity of the benchmarking process. It was not shortlisted for these reasons.

106. The five remaining factors, **household residual income, key stage 4 – maintained school status, Income deprivation affecting children index (IDACI), socio-economic background and Tracking underrepresentation of areas (TUNDRA) quintile** all had estimated differences that were among the smallest across the longlisted factors. This means that they represent a weaker fit with the benchmarking principles because they are less correlated with the outcome.[34] The factors are all also considered to overlap with other, higher priority, student characteristics already included in the shortlist. Taken together, we considered that this meant that inclusion of the factors would risk the statistical integrity of the benchmarking process. The factors were not shortlisted for these reasons.

International students

107. Throughout the initial modelling, international students were identified through the home region factor, which was not shortlisted. These students are included in the coverage of all indicators except the progression indicator. For indicators other than the progression indicators, there are three factors that have been shortlisted which would mean that these students would be isolated into their own group. These are the ABCS quintile, entry qualifications, and ethnicity. In all cases, these indicators are most applicable to UK domiciled students only. In order to benchmark international students, we have created explicit additional groupings within those factors to separate these students from students that have been otherwise categorised as ‘Unknown’ in the initial modelling. English deprivation quintile (applicable only to English-domiciled students) was also shortlisted for continuation, completion and progression measures as discussed above.

[34] For both socio-economic background and TUNDRA quintile we also considered results from bespoke statistical modelling. This did not change our view of the inclusion of these factors once other factors are removed from the models to reach a shortlist.

Findings relevant to determining the proposed benchmarking factors

108. The findings that follow in this section reflect the judgements we have made about which of the shortlisted factors to include in our proposed benchmarking factors. They describe the careful consideration we have given to each factor, following the method described in paragraphs 39 to 42.

109. We assessed modelling based on the shortlisted factors and identified that all shortlisted factors retained statistically significant estimated differences. This is illustrated in Table 9 below. However, the shortlist contained too many factors to include all of them, so further prioritisation is required to maintain statistical integrity of the benchmarking process. This is because it is not considered feasible that we include all of the shortlisted factors: to do so would risk our ability to maintain statistical integrity of the benchmarking process, by creating benchmarking groups defined at a very granular level and so likely each containing very small numbers of students.[35]

110. Consequently, the section that follows Table 9 then considers the relative priority of each factor, for each indicator. Throughout the commentary we provide in this section of the report, the descriptions we give are summative, and should not be considered as exhaustive of our considerations of each factor, nor determinative of the proposals we have made through our consultations.

111. The final section in this set of findings describes an overview of our expert judgement in selecting the proposed factors (those discussed in Proposal 10 of the indicators consultation). In describing those expert judgements, we discuss our assessment of the appropriate balance of the relative priority for a factor with the statistical integrity of the benchmarking method as a whole.

Illustrating the statistical modelling of the shortlisted factors

112. Table 9 takes the same approach as Table 8 above, as outlined in paragraph 51. It presents the largest estimated differences (for the attribute with the largest significant difference in each case) for each of the shortlisted factors by indicator and mode of study. The table only shows estimated differences that are statistically significant at the 95 percent level, and excludes 'unknown' and similar attributes. Note that many factors had multiple significant differences, and we considered differences across all of the attributes, as well as the numbers of students associated with each attribute, when making our judgements.

113. Our analysis showed that there was reasonable consistency in the relative importance of factors across the NSS scales that are used to construct the student experience indicators. We considered that there was not sufficient evidence of variation to warrant selecting different factors for the different student experience indicators. Hence, for brevity, in this report the data relating to student experience indicators are represented by results from the 'teaching on my course' scale. The full modelling results are available for the five scales that

[35] To mitigate this risk, we considered whether any attributes of our shortlisted factors could be grouped further to facilitate inclusion of all of the shortlist. The modelling results did not suggest that we could do this in a coherent way, in particular for subject and entry qualifications. Both factors, given their granularity, represent the highest risk to the statistical integrity because they have the most attributes.

are proposed for use in the TEF consultation including the 'teaching on my course' scale in the accompanying results workbooks.

Table 9: Maximum estimated differences for the shortlisted factors, by indicator and mode of study

Key: Cells marked with grey shading identify factors not shortlisted for a particular indicator or mode. Cells showing '-' indicate no significant differences for that factor. Estimated differences in brackets correspond to populations of fewer than 5,000 students, where the difference shown in the same cell without brackets is the largest estimated difference for populations of at least 5,000.

Maximum estimated differences (percentage points, significant at the 95 per cent level only)	**Continuation**		**Completion (cohort tracking)**		**Progression**		**Student experience (teaching on my course scale)**	
Factor	Full-time	Part-time	Full-time	Part-time	Full-time	Part-time	Full-time	Part-time
Age on entry	1.5	2.0	2.2	1.6	5.5	3.6	6.1	3.3
Associations between characteristics of students (ABCS) quintile	3.9	12.8						
Deprivation quintile (IMD)	1.1	1.8	2.2	2.9	3.0	2.5		
Disability	0.9	2.5	2.0	3.4	1.7	2.5	1.5	2.2
Eligibility for free school meals	2.0	8.7	3.1	10.2				
Entry qualifications	13.8	16.0 (20.9)	19.6	14.7	13.2	5.7	3.5	-
Ethnicity	0.5	1.5	0.5	2.9	3.7	4.8	4.6	- (1.9)
Expected course length in years and short course		5.9	8.0	12.2				
First degree course with integrated foundation year	5.5		10.6					
Geography of employment quintiles					7.8	4.5		
Level of study	4.5	4.9 (5.2)	4.8	1.1 (9.0)	5.0	1.4	4.3	-
Location of study	2.7							

Maximum estimated differences (percentage points, significant at the 95 per cent level only)	Continuation		Completion (cohort tracking)		Progression		Student experience (teaching on my course scale)	
Sex	1.7	2.6	3.2	2.6	1.9	3.7	1.1	-
Subject	2.2	12.5 (21.6)	4.1	11.2 (14.2)	27.1	6.9 (24.5)	9.1	7.0
Year					2.5	-	4.3	1.3

Assessment of the priority of the shortlisted factors

114. In making the following judgements about which of the shortlisted factors have been prioritised for consideration as proposed benchmarking factors, we carefully considered each factor following the method described in paragraphs 39 to 42. We provide a commentary in paragraphs 115 to 135 which summarises our judgement of the relative priority of retaining each of the shortlisted factors for the purposes of formulating a viable set of proposed benchmarking factors. These judgements build on the considerations (through paragraphs 56 to 107 above and not repeated here) of each factor when it was considered at the shortlisting stage.

Subject of study

115. Having been shortlisted for all indicators for the reasons described in paragraphs 56 to 57, subject of study was considered among the highest priority to **retain for all indicators**, for both modes of study. Modelling of the shortlisted factors shows that across the indicators, most attributes had a statistically significant estimated difference. The sizes of those estimated differences differ by indicator and mode, but there was no meaningful evidence to suggest that the groupings of the attributes for this factor were inappropriate or could be grouped further in a coherent way. Our consideration of this factor as one of the highest priorities to retain relies on these modelling findings, together with the compelling reasons for shortlisting this factor described in paragraph 56.

Entry qualifications

116. Having been shortlisted for all indicators for the reasons described in paragraphs 58 to 61, entry qualifications were considered among the highest priority to **retain for continuation, completion, and progression indicators**, for both modes of study. The shortlist modelling shows that this factor is among the factors with the largest estimated differences in outcomes across its attributes. The sizes of those estimated differences differ by indicator and mode, but there was no meaningful evidence to suggest that the groupings of the attributes for this factor established at the shortlisting stage were inappropriate, or could otherwise be grouped in a more coherent way. Our consideration of this factor as one of the highest priorities to retain for continuation, completion, and progression indicators relies on these modelling

findings, together with the compelling reasons for shortlisting this factor described in paragraph 58 and 59.

117. For the student experience indicators, our modelling did not show particularly large estimated differences between attributes. Together with the concern described in paragraph 59, that it may not be appropriate for benchmarking to account for an apparent association between entry qualifications and responses to the NSS in the final year of study, this has led us to lower the priority of this factor and **not retain this factor for student experience indicators**, for either mode of study.

Level of study

118. Having been shortlisted for all indicators for the reasons described in paragraphs 62 to 64, level of study was considered among the highest priority to **retain for all indicators**, for both modes of study. Modelling of the shortlisted factors shows that across all indicators, there were attributes with statistically significant estimated differences. For the full-time continuation and completion indicators, we found estimated differences in our modelling between Level 4 and Level 5 other undergraduate provision that were large enough to justify differentiating between these levels for those indicators. Our consideration of this factor as one of the highest priorities to retain relies on these modelling findings, together with the compelling reasons for shortlisting this factor described in paragraph 62.

Year

119. Having been shortlisted for the progression and student experience indicators for the reasons described in paragraphs 65 to 66, year was considered high priority to **retain for progression and student experience indicators**, for both modes of study. The shortlist modelling results support the initial modelling in finding correlations that could relate to the impact of the pandemic. Our consideration of this factor as a high priority to retain for these indicators relies on these modelling findings, together with the reasons for shortlisting this factor for these indicators described in paragraph 65.

First degree course with integrated foundation year

120. Having been offered benefit of the doubt in being shortlisted for the full-time continuation and completion indicators for the reasons described in paragraphs 67 to 69, first degree courses with integrated foundation year were considered lower priority and **not retained for continuation and completion indicators.** The shortlist modelling continues to show estimated differences for the attributes of this factor.

121. We considered the impact of this factor overlapping with that for entry qualifications, which (as described in paragraph 116) we have identified as a high priority for retaining. While the shortlist modelling identifies an estimated difference for foundation year courses over and above the influence of entry qualifications (this factor is held in the modelling when calculating the estimated differences for foundation year courses), the overlap between the two factors led us to choose one to prioritise. This is because prioritising both would potentially take limited additional account of student preparedness for starting higher education. At the same time, it would increase the risk to the statistical integrity of the benchmarking approach by increasing the number of benchmarking groups and hence the likelihood that these are sparsely populated. We considered that it would not be desirable to include the foundation year factor at the expense of entry qualifications, or of other high

priority factors which would need to be deprioritised in order to preserve the statistical integrity of the method.

122. Our preference was to deprioritise the foundation years factor because it is unclear that the remaining effect for foundation years is still accounting for preparedness for starting higher education, as opposed to acting as a proxy for provider performance. We preferred to prioritise the entry qualifications factor because it has a greater correlation than that of the foundation year factor, and wider applicability across the student population considered by the indicators. We have therefore chosen to lower the priority of the foundation year factor and **not retain this factor for full-time continuation and completion indicators**.

Expected course length in years (and short courses)

123. Having been shortlisted for the full-time completion indicator for the reasons described in paragraphs 70 to 71, the expected course length factor was considered high priority to **retain for the full-time completion indicator.** The shortlist modelling continues to show large estimated differences for the three course length groupings defined as attributes of this factor for this indicator. Our consideration of this factor as a high priority to retain for this indicator relies on these modelling findings, together with the reasons for shortlisting this factor for this indicator described in paragraph 70.

124. Having also been shortlisted for the part-time continuation and completion indicators for the same reasons (described in paragraphs 70 to 71), a short course factor derived from expected course length (courses lasting less than 24 weeks) was considered high priority to **retain for the part-time continuation and completion indicators.** The shortlist modelling continues to show large estimated differences for the attributes of the factor defined on this basis. Our consideration of this factor as a high priority to retain for these indicators relies on these modelling findings, together with the reasons for shortlisting this factor for this indicator described in paragraph 70.

Study location type (identifying distance learning and local students) and commuting student

125. Having been shortlisted for full-time continuation indicator for the reasons described in paragraphs 74 to 75, the study location type factor was considered lower priority and **not retained for the full-time continuation indicator.** The shortlist modelling results support the initial modelling in finding smaller estimated differences for the attributes of this factor.

126. We considered the impact of the factor already being indirectly considered through its inclusion in the definition of the ABCS quintiles for continuation outcomes, which (as described in paragraph 128) we have identified as a high priority for retaining. While the modelling identifies an estimated difference over and above the influence of the ABCS quintiles (this factor is held constant in the modelling when calculating the estimated differences), the overlap between the two led us to choose one to prioritise. This is because prioritising both would potentially take limited additional account of study locations, while at the same time it would increase the risk to the statistical integrity of the benchmarking approach by increasing the number of benchmarking groups and hence the likelihood that these are sparsely populated. We consider that it would not be desirable to include the study location type factor at the expense of the ABCS quintiles, or of other high priority factors which would need to be deprioritised in order to preserve the statistical integrity of the

method. We have therefore chosen to lower the priority of the study location type factor and **not retain this factor for the full-time continuation indicator**.

Geography of employment quintile

127. Having been shortlisted for the progression indicators for the reasons described in paragraphs 77 to 80, geography of employment quintile was considered high priority to **retain for progression indicators**, for both modes of study. Modelling of the shortlisted factors shows that across all indicators, there were attributes with statistically significant estimated differences. Our consideration of this factor as a high priority to retain for the progression indicators relies on these modelling findings, together with the compelling reasons for shortlisting this factor for this indicator described in paragraph 77.

Associations between characteristics of students (ABCS) quintile

128. Having been shortlisted for the continuation indicators for the reasons described in paragraphs 93 to 96, ABCS quintile was considered among the highest priority to **retain for continuation indicators**, for both modes of study. Modelling of the shortlisted factors shows that the ABCS quintiles had relatively large estimated differences for full-time and very large estimated differences for part-time. It is important to note that the modelling results reflect the remaining differences between attributes with other factors in the models held constant, including student characteristics that were used in the derivation of the ABCS. If some, or all, of these factors were removed from the models, we would expect to see the estimated differences become larger for the ABCS and our bespoke modelling confirmed this.

129. Our consideration of this factor as one of the highest priorities to retain for the continuation indicator relies on these modelling findings, together with the compelling reasons for shortlisting this factor for this indicator described in paragraph 94.

Eligibility for free school meals

130. Having been offered benefit of the doubt in being shortlisted for the continuation and completion indicators for the reasons described in paragraphs 100 to 103, eligibility for free school meals was considered lower priority and **not retained for continuation and completion indicators.**

131. For part-time continuation indicators, and for completion indicators in both modes of study, the shortlist modelling shows statistically significant estimated differences. Our consideration of this factor as lower priority therefore relies on the concerns raised about the partiality of the coverage of this factor for these indicators, as described in paragraph 101. We considered that the number of students included in these indicators and classified as eligible for free school meals is understated due to the limitations of the data available for linking (with too many students classified with unknown eligibility). This would mean that the factor would potentially take limited additional account of student backgrounds, and at the same time it would increase the risk to the statistical integrity of the benchmarking approach by increasing the likelihood of sparsely populated benchmarking groups. We have therefore chosen to lower the priority of the free school meals factor and **not retain this factor for the part-time continuation indicators, or for completion indicators.**

132. As also described in paragraph 101, we consider that the full-time continuation indicator is less impacted by the concerns about the partiality of the coverage of this factor. However,

while the shortlist modelling also showed statistically significant estimated differences for the attributes of this factor for this indicator, these are relatively small. Furthermore, we considered the impact of the factor already being indirectly considered through its inclusion in the definition of the ABCS quintiles for continuation outcomes, which (as described in paragraph 128) we have identified as a high priority for retaining. While the modelling identifies an estimated difference over and above the influence of the ABCS quintiles (this factor is held constant in the modelling when calculating the estimated differences), the overlap between the two led us to choose one to prioritise. This is because prioritising both would potentially take limited additional account of student backgrounds, while at the same time it would increase the risk to the statistical integrity of the benchmarking approach by increasing the number of benchmarking groups and hence the likelihood that these are sparsely populated. We consider that it would not be desirable to include the free school meals factor at the expense of the ABCS quintiles, or of other high priority factors which would need to be deprioritised in order to preserve the statistical integrity of the method. We have therefore chosen to lower the priority of the free school meals factor and **not retain this factor for the full-time continuation indicator**.

Other student characteristics: Age, disability, ethnicity, IMD, sex

133. Having been shortlisted for all indicators for the reasons described in paragraphs 85 to 89, the other student characteristic factors of age, disability, ethnicity, IMD quintile and sex were considered high priority and **to retain for all indicators except continuation indicators.** They were considered lower priority and **not retained for continuation indicators.** The shortlist modelling show statistically significant estimated differences for the attributes of each factor, which are similar across all indicators.

134. For continuation indicators, we considered the impact of these factors all already being indirectly considered through their inclusion in the definition of the ABCS quintiles for continuation outcomes, which (as described in paragraph 128) we have identified as a high priority for retaining. While the modelling identifies estimated differences over and above the influence of the ABCS quintiles (this factor is held constant in the modelling when calculating the estimated differences), the overlaps between them led us to choose one to prioritise. This is because prioritising ABCS quintiles as well as these characteristics would potentially take limited additional account of student backgrounds. At the same time, it would significantly increase the risk to the statistical integrity of the benchmarking approach by increasing the number of benchmarking groups and hence the likelihood that many of these are sparsely populated. We considered that it would not be desirable to include these factors at the expense of the ABCS quintiles, or of other high priority factors which would need to be deprioritised in order to preserve the statistical integrity of the method. We also considered that it is not possible or desirable to make a meaningful choice between the student characteristic factors of age, disability, ethnicity, IMD quintile and sex, all of which have similar estimated differences and coherence with policy objectives. We therefore have chosen to lower the priority of these factors and **not retain this factor for continuation indicators**.

135. For all other indicators, our consideration of these factors as high priorities to retain relies on the shortlist modelling findings, together with the compelling reasons for shortlisting them described in paragraph 85. It also acknowledges the current absence of ABCS quintiles for the student outcomes and experiences to which the other indicators refer. We have therefore

considered it high priority to **retain age on entry, disability, ethnicity, deprivation quintile (IMD) and sex for all other indicators (except continuation)**, for both modes of study.

Assessment of the prioritised factors

136. As noted at paragraph 109, it is not considered feasible that we include all of the shortlisted factors: to do so would risk our ability to maintain statistical integrity of the benchmarking process, by creating a large number of benchmarking groups defined at a very granular level, making it likely that these groups were sparsely populated by small numbers of students.

137. Through the discussion in paragraphs 114 to 135 above, we have established a set of factors that have been prioritised for each indicator. In this section, we present analysis of these prioritised factors, which informed our final judgements in selecting the proposed factors (those discussed in Proposal 10 of the indicators consultation). The analysis looks at each indicator and mode of study, and includes information intended to support an informed view of the appropriate balance of the relative priority for a factor with the statistical integrity of the benchmarking method as a whole:

    a. The theoretical number of distinct benchmarking groups. This is presented in Table 10.

    b. The actual number of distinct benchmarking groups with at least one student. This is presented in Table 10.

    c. The number of populated benchmarking groups that contain relatively few students (i.e. those with five or fewer students, or 20 or fewer students). This is presented in Table 10.

    d. The actual contribution to benchmark for each provider. This additionally includes the apprenticeship mode of study, where the factors prioritised for the equivalent part-time indicator have been used.[36] This is presented in Tables 11 and 12. The underlying methodology to creating this statistic is described in the statistical methods document.[37]

138. For the student outcomes indicators, we include two separate analyses: one with the prioritised student characteristics and one without. We do this because in developing our proposals for the selection of benchmarking factors, we took the view that it is not possible for us to include all of the student characteristic factors which are correlated with our indicators, in addition to other factors we considered a high priority. The two parts of the analysis consider the two extremes of the approaches that we considered in relation to the

[36] As noted in paragraph 21, the potential for conducting the appropriate statistical modelling is more limited on account of the more limited spread and characteristics of apprenticeship students across the sector. When considered at the level of detail necessary within the models, there are currently insufficient student numbers for those models to be robust (or, in technical terms, to converge). The benchmarking factors that we have proposed in the indicators consultation in most cases involve using the same benchmarking factors for indicators constructed for apprenticeship students as for part-time students. The proposed benchmarking factors deviate between apprenticeship and part-time indicators for continuation and completion indicators for the expected course length factor. This is because very few apprenticeship students have an expected course length of less than a year.

[37] See supporting documents at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/outcome-and-experience-data/.

inclusion of the prioritised student characteristics. That is, we examine two scenarios based on including all of them or including none of them.

139. We also considered the potential to select one or a subset of the student characteristic factors of age, disability, ethnicity, IMD quintile and sex. As described in paragraph 134, we consider that it is not possible or desirable to make a meaningful choice between these factors, all of which have similar estimated differences and coherence with policy objectives. The analysis includes the continuation indicator, even though the high priority given to the inclusion of ABCS quintile as a factor means that we have considered the student characteristic factors a lower priority, as stated in paragraph 134. The inclusion of the continuation indicator in the analysis aims to demonstrate the impact of including further factors on the statistical integrity of the benchmarking process.

Table 10: Number of attributes, combinations, and small groups from the remaining prioritised factors

Factor	Continuation		Completion (cohort tracking)		Progression		Student experience	
	Full-time	Part-time	Full-time	Part-time	Full-time	Part-time	Full-time	Part-time
Age on entry	3	3	3	3	3	3	3	3
Associations between characteristics of students (ABCS) quintile	6	6						
Deprivation quintile (IMD)	3	3	3	3	3	3		
Disability	2	2	2	2	2	2	2	2
Entry qualifications	11	5	11	5	10*	5		
Ethnicity	3	3	3	3	2*	2*	6	6
Expected course length in years			3					
Short course		2		2				
Geography of employment quintiles					3	3		
Level of study (3 when other undergraduate not split)	4	3	4	3	3	3	3	3
Subject (34 for CAH2, 21 for CAH1, 10 for broad grouping)	21	21	21	21	34	10	34	10
Year of indicator					2	2	4	4
Sex	2	2	2	2	2	2	2	
Without prioritised student characteristics								
Number of possible combinations	5,540	3,780	2,770	630	6,120	900		
Number of populated combinations	4,250	1,520	1,820	350	4,570	630		
% of students in very small groups (1 to 5)	0.1%	0.3%	<0.1%	<0.1%	0.8%	0.7%		
% of students in small groups (1 to 20)	0.6%	1.3%	0.3%	0.1%	3.2%	5.4%		
With prioritised student characteristics								
Number of possible combinations	598,750	408,240	299,380	68,040	440,640	64,800	29,370	4,320
Number of populated combinations	64,510	24,820	37,480	19,260	63,750	10,040	14,600	1,830
% of students in very small groups (1 to 5)	4.0%	9.8%	2.1%	4.5%	20.4%	38.1%	0.5%	1.1%
% of students in small groups (1 to 20)	11.5%	25.1%	5.9%	12.0%	45.0%	73.0%	2.2%	4.8%

* Progression is a UK-only indicator, so non-UK groups are ignored

140. In Table 10 above, we can see how the number of possible and populated benchmarking factor combinations vary, as well as how the proportion of students in small groups of up to five or 20 students vary when comparing the inclusion of prioritised student characteristics or not, across indicators.

141. Smaller benchmarking groups increase the statistical uncertainty associated with the benchmarking (as reflected by our calculated standard deviations)[38] and when many benchmarking groups are populated by only one or two students, the sector performance calculated for those groups will tend to a small range of values. For example, if in the benchmarking group contains a single student, it can only result in either 0 per cent or 100 per cent. If it refers to only two students, the average can only be 0 per cent, 50 per cent or 100 per cent. Benchmarking groups with performance that include large numbers of 0 per cent and 100 per cent values can lead to an ineffectual weighting which will skew the resulting benchmark. Therefore, having too many students in small benchmarking groups could limit our ability to assess performance relative to the benchmark.

142. Examining the results in Table 10, it can be seen that for continuation, completion and progression indicators, the inclusion of all student characteristics factors results in very large numbers of combinations, and moderate to high proportions of students in small groups. This suggests that we cannot include all of the student characteristics as well as the other factors that we have deemed high priority.

143. To better understand how these small benchmarking groups impact the benchmarking process, we have calculated the contribution of providers towards their own benchmarks. This statistic indicates the influence of the provider's own students on the sector averages that informs the calculation of the provider's benchmark, and helps us understand the risk of self-benchmarking (which can arise when the characteristics of students at the provider in question do not frequently occur among students in the wider sector, and the provider's own students would be making a substantial contribution to its benchmark). Where this happens the utility of the benchmark is undermined. The following tables show the proportion of providers that contribute more than 5 per cent, 20 per cent, or 50 per cent towards their own benchmarks.[39] Table 11 shows results without the prioritised student characteristics (as outlined above). Table 12 shows results with all the prioritised characteristics included.

[38] These standard deviations are described in further detail in the statistical methods document.

[39] Only providers with a denominator population of at least 23 students are considered for these statistics. The part-time undergraduate with postgraduate components mode and level combination is not included in these tables due to insufficient numbers of providers meeting this denominator restriction.

Table 11: Providers' contributions to their own benchmarks (without prioritised student characteristics)

Percent of providers with at least (X%) contribution to their own benchmark	Continuation				Completion (cohort tracking)				Progression				Student experience (teaching on my course scale)			
Mode and level	Number of providers	> 5%	>20%	>50%	Number of providers	> 5%	>20%	>50%	Number of providers	> 5%	>20%	>50%	Number of providers	> 5%	>20%	>50%
Full-time first degree	288	10%	0%	0%	259	12%	1%	1%	246	22%	0%	0%				
Full-time other undergraduate	293	38%	9%	2%	277	35%	6%	2%	222	84%	12%	1%				
Full-time undergraduate with postgraduate components	93	75%	10%	3%	80	63%	10%	4%	71	99%	28%	4%				
Part-time first degree	136	33%	5%	1%	139	15%	4%	2%	80	39%	1%	1%				
Part-time other undergraduate	246	28%	6%	1%	251	12%	3%	0%	156	29%	5%	0%				
Apprenticeship (all undergraduates)	204	16%	2%	0%	116	17%	4%	0%	20	100%	65%	10%				

Table 12: Providers' contributions to their own benchmarks (with prioritised student characteristics)

Percent of providers with at least (X%) contribution to their own benchmark	Continuation				Completion (cohort tracking)				Progression				Student experience (teaching on my course scale)			
Mode and level	Number of providers	> 5%	>20%	>50%	Number of providers	> 5%	>20%	>50%	Number of providers	> 5%	>20%	>50%	Number of providers	> 5%	>20%	>50%
Full-time first degree	288	48%	2%	0%	259	47%	7%	1%	246	100%	35%	2%	242	21%	2%	0%
Full-time other undergraduate	293	98%	33%	7%	277	92%	26%	5%	222	100%	99%	28%	244	73%	13%	2%
Full-time undergraduate with postgraduate components	93	100%	51%	11%	80	95%	24%	8%	71	100%	92%	27%	83	92%	23%	6%
Part-time first degree	136	95%	33%	3%	139	52%	8%	3%	80	100%	85%	4%	77	35%	4%	1%
Part-time other undergraduate	246	79%	17%	4%	251	35%	8%	0%	156	100%	71%	6%	106	75%	17%	0%
Apprenticeship (all undergraduates)	204	71%	8%	1%	116	68%	10%	1%	20	100%	100%	80%	73	82%	10%	1%

144. Comparing the figures in Table 11 and Table 12, we can see the impact on the amount of self-benchmarking of benchmarking by the prioritised student characteristic factors in addition to the other prioritised factors.

145. For the continuation indicator, our proposed factors do not include individual student characteristics, as discussed above in paragraph 134. The associated number of theoretical and populated benchmarking groups, number of students in small groups, available in the 'without prioritised student characteristics' section of Table 10, and the assessments of provider contributions to benchmark, available in Table 11, indicate that our proposal does not result in too many small groups or too much self-benchmarking. The equivalent figures shown in Table 10 which include the additional student characteristics that were not prioritised show that there is a significant increase in the level of self-benchmarking, which is undesirable for the reasons discussed in paragraphs 141 to 143. We therefore believe that the judgements made about the prioritisation of the shortlisted factors are appropriate and proportionate, given the modelling results and the inclusion of the intersectional ABCS factor.

146. For the other student outcomes indicators (completion and progression), the numbers in Table 12 demonstrate that the inclusion of all the prioritised factors is not possible without substantial amounts of self-benchmarking, which we judge would risk the statistical integrity of the benchmarking process. The equivalent figures in Table 11, with the prioritised student characteristics not included, represent a more manageable risk of self-benchmarking. For these indicators, the prioritised individual student characteristics (age, disability, ethnicity, sex, and IMD) all had small to medium estimated differences in our shortlist modelling, suggesting that no one was particularly higher priority than another. As it is not feasible to include them all as individual factors, and we do not wish to select some but not all of them (as discussed in paragraph 138 above), we plan to develop ABCS measures for these indicators and propose to implement these as benchmarking factors for the completion and progression indicators. This will ensure that we can account for student characteristics in the benchmarking, without compromising the statistical integrity of the process or sacrificing the inclusion of any of the other prioritised factors.

147. We have selected the same benchmarking factors across all of the scales derived from the NSS for the student experience indicators, because our analysis showed significant consistency of the estimated differences and correlations of factors across the scales. For these indicators, all the prioritised individual student characteristics showed significant estimated differences between attributes (except sex for the part-time indicator). Again, we do not wish to select to some but not all of these, so we prioritise including all of them. Despite having four (full-time) or three (part-time, excluding sex) factors as well as the other prioritised factors, we did not find excessive numbers of students falling into small benchmarking groups (see Table 10). This is due to the omission of entry qualifications as a factor which allows us to use more benchmarking factors overall. Furthermore, the figures in Table 12 are broadly in line with the numbers in Table 11 relating to the student outcomes indicators without the student characteristics. We judge that these figures represent a manageable risk of self-benchmarking that is justified by the inclusion of the benchmarking factors we have proposed. Based on these findings, and in recognition of the current ongoing

review of the NSS[40], we do not expect to develop ABCS measures for these indicators in the immediate future.

148. We judge that our proposed factors represent a range of factors which strike an appropriate balance between consistency with our policy objectives (we consider that the list of proposed factors is broad enough to account for many of the characteristics and course structures correlated with differences in student outcomes and experiences), and the statistical integrity of the benchmarking approach, to support a benchmarking approach that is fit for purpose. While we have needed to prioritise the selection of some factors over others, we take the view that we have taken appropriate steps to minimise the number of students likely to be in small groups, and the number of providers with substantial contributions to their own benchmarks. The proposed factors for each indicator are shown in Tables 12 to 15 of the indicators consultation.

[40] For more information, see www.officeforstudents.org.uk/advice-and-guidance/student-information-and-data/national-student-survey-nss/review-of-the-nss/.

Annex A: Statistical modelling methodology

149. Datasets are created for each indicator as defined in 'Instructions for rebuilding OfS datasets'.[41] For modelling purposes, we are interested in the relationship between the various factors and the outcome (i.e. those which count positively towards the numerator of the indicator). In this section we will use continuation as an example. In this case, a positive outcome corresponds to a student either qualifying or actively studying in the following year (two years for part-time).

150. A binary logistic regression model is fitted to the indicator data. The dependent variable is the outcome in question; the factors are included as independent variables. Teaching provider[42] is included as a fixed effect to account for the variation in outcomes between providers that is not due to student characteristics. The generalised formula for the model is as follows:

Equation 1: Logistic regression of student characteristics of continuation rates

$$\boldsymbol{Student\ outcome_i \sim Bernoulli(N, \pi_i)}$$

where N is the weighted total of entrants in the modelling population and π_i is the probability of continuation for individual i.

$$logit(\pi_i) = \beta_0 + \beta_1(Factor\ 1)_i + \beta_2(Factor\ 2)_i \ldots + \beta_n(Factor\ n)_i + \beta_p Provider_i$$

where the β terms are vectors of varying sizes relating to each factor, and $Provider$ is a factor representing the individual's teaching provider.

151. The outputs of this model include parameter estimates for each of the attributes of each factor, and p-values showing the statistical significance of each parameter estimate.[43] We also calculate odds ratios and 'estimated differences in outcomes' (outlined in the following section). These help to confirm whether the results vary across attributes and if they can be said to be non-trivial.

152. If any of the attributes show a statistically significant difference (as indicated by p-values) from the reference group, then we can be confident that the factor is correlated with the outcome, taking all other factors into consideration. By default, the reference group is chosen to be the largest category within the characteristic in question, unless this is the 'unknown' group. However, in some cases where categories are of a similar size, we have chosen the category with either highest or lowest performance, to maximise the estimated differences that we observe.

[41] See supporting documents at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/outcome-and-experience-data/.

[42] The teaching provider during the academic year for which the student outcome is associated is used.

[43] In rare cases, the model cannot calculate parameters for some attributes. This can be due to separation of outcomes, where another factor in the model completely separates the attribute in question from the reference category. Where this occurs, we have treated results with caution.

Calculating estimated differences in outcomes between attributes

153. This section uses continuation as an example to aid understanding, but the method is applicable to all indicators. For example, instead of ‘continuation rate’ the outcome may be ‘proportion of students progressing to highly skilled employment or further study’ for the progression indicator.

154. Model predictions are obtained for individuals where the characteristic of interest (e.g. their age group) is forced to be equal to the reference group (e.g. ‘Under 21’) when the observed actual value is not (e.g. the student is actually aged 21 to 30). This then yields an estimate of the likelihood of continuation for each student aged 21 to 30 on entry to their course if they were instead aged under 21 and all other characteristics remained unchanged.

155. An average of this group’s continuation rates is then taken to calculate the ‘model estimated continuation rate’. 95 per cent confidence intervals around the model-estimated continuation rates are derived from the standard errors associated with the fixed effects for each attribute.

156. This method is mathematically equivalent to transforming the relevant parameter estimate from the statistical model into a ‘marginal effect’ in percentage points, relative to the reference category. With categorical variables, this result is more straightforward to interpret than an ‘average marginal effect’; the ‘average’ student will invariably sit between categories recorded in the data, which makes the result hypothetical and harder to interpret.

157. These model predictions cannot truly capture the continuation rate of a given student if they were in the reference group while their other characteristics remained unchanged, because this is hypothetical. For this to hold, we must assume that the parameter estimates of the fixed effects represent causal relationships: for example, if a chemistry student studied business and management instead, the estimate of the model coefficient for studying chemistry is how much their likelihood of continuation would change (after applying a logarithmic transformation).

158. The difference between the model-estimated continuation rates if entrants belonged to the reference group and the actual continuation rates for each group then indicate an ‘estimated difference in outcomes’ of belonging to their group instead of the reference group. When this value is negative (the modelled rate is higher than the actual rate), it means that on average, the model predicts that these students would be more likely to have a positive continuation outcome if they belonged to the reference group instead, and vice versa. This is the modelled effect of changing this one characteristic, holding all other characteristics the same.

159. An advantage of using this method is that it provides a clear and intuitive way of understanding the relationship between each attribute and the outcome. The confidence ranges also emphasise where an estimated difference is not statistically significantly different from zero. It is possible with this method to compare estimated differences between different factors, to get an idea of the typical magnitude of differences. This is useful when considering the policy aspects of factors, as it allows non-expert readers to understand the relative importance of different factors. Alternative methods were considered.[44]

Confidence intervals

160. Confidence intervals indicate how much (observed) uncertainty there is around a given statistic; we would expect the true value of the statistic to lie between the intervals 95 per cent of the time, given the data in front of us. In other words, we are 95 per cent 'confident' that the true value lies between the intervals.

161. It is important to note that the 95 per cent significance level was primarily chosen to be illustrative of the observable statistical uncertainty. It also provides a tolerance of 'Type II' errors that suits our uses on this occasion, based on our expert judgement.[45]

162. For our estimated differences in outcomes, 95 per cent confidence intervals are obtained by first adjusting the predicted probability for each student according to the standard error around the parameter estimate for the characteristic that has been set to the reference category, as described in the previous section. For example, the predicted probabilities for male students 'as if they were female' are adjusted by $\pm 1.96 \times \text{s.e.}(\hat{\beta}_{Sex,Male})$, where $\text{s.e.}(\hat{\beta}_{Sex,Male})$ is the standard error for the estimate of the fixed effect of being male, relative to female.

163. As before, averages are then taken at the sector level to obtain model estimated rates of positive outcomes where the characteristic is set to the reference group, but this time each of the predicted probabilities have been adjusted by $\pm 1.96 \times \text{s.e.}(\hat{\beta})$. Then the differences between the actual rate and the upper and lower estimates of the rate (where the characteristic was set to the reference group) indicate the intervals between which we are 95 per cent confident the true difference lies, after controlling for underlying differences in the other factors included in the model.

[44] We also examined likelihood ratio tests to identify significant attributes. Likelihood ratio tests may be carried out between nested models. Two models are nested when one can be derived from the other by removing one or more parameters. The likelihood ratio test is carried out by comparing the '-2 log likelihood' values from the two models. A corresponding significance (p-value) can then be calculated from this number and the difference in degrees of freedom between the two models. We compared a 'full' model, using all candidate factors, with a 'reduced' model, with one factor removed. The comparison between these two nested models then shows whether this factor significantly improves model fit.

[45] In statistical hypothesis testing, a Type II error occurs when one accepts a null hypothesis that is actually false. For example, under the null hypothesis that a given difference in continuation rates is equal to zero, we would be making a Type II error if we concluded, given the data in front of us, that the difference in continuation rates was equal to zero, but the true value was in fact different from zero. Selecting a higher level of statistical significance means a Type II error is less likely to occur.

Example interpretation of estimated differences

164. Having run the model with all possible factors included, and carried out the above process, we can examine the results for a single factor, 'Year', in Table B1.

Table B1: Example of estimated differences between attributes for Year

Attribute	p-value	Observed rate	Predicted rate if each group were in the reference group	Difference between modelled and observed rates (with 95% confidence range)
2014	0.045	90.9%	90.8%	-0.1% (-0.2% to 0.0%)
2015 (reference)	-	90.8%	-	-
2016	0.0255	90.9%	90.8%	-0.1% (-0.2% to 0.0%)
2017	<.0001	90.5%	90.7%	0.3% (0.1% to 0.3%)

165. In this table, 2015 is the reference group. The reference group is normally chosen to be the largest group of interest (i.e. excluding 'unknown' groups). An observed rate can be calculated for this group, and all model estimates are calculated relative to this group. The modelled rate shows the predicted rate if the population of each of the other groups was changed to be in the reference group, and all other factors held equal.

166. It can be seen that for the year 2017, the p-value is significant and we can be confident that the difference between the modelled and observed rates is not zero; hence the principle of correlation between the factor and the outcome is satisfied. However, the difference between the attributes (0.3 per cent for 2017 compared to 2015) is arguably trivial; this will be compared to the estimated differences for other factors when considering which factors to eventually use for benchmarking. For example, if other factors had differences several times larger, then they might be chosen in preference.

Interactions

167. The results of the statistical modelling tell us which factors are significantly correlated with the outcome and which are not. In the description that follows, these factors are referred to respectively as primary and secondary.

168. Before reaching our shortlist, as described in paragraph 36, we want to identify instances where the effect of belonging to a particular group is only significant when we divide that group by a second factor. As an example, say there was no significant relationship between disability and continuation, but when taking a primary factor, sex, into account a significant relationship was revealed between male disabled students and propensity to continue. In this example, the secondary factor, disability, could still be a suitable benchmarking factor when used in addition to sex.

169. To avoid overlooking any such factors we model every two-way interaction between the primary factors and each secondary factor. The resulting number of combinations would be impractical to examine manually. Hence, each combination is compared against the non-

interacted equivalent using AIC46. If the AIC term is smaller for the model with interactions, this shows that the interactions are adding information to the model. The parameter estimates of these cases are then inspected manually, to see if any of the attributes of the interaction terms have significantly different outcomes to the reference groups.

170. To illustrate this process, consider an example: if factors A, B, and C are primary factors, and factor M is a secondary factor, the following models are run.

1. $logit(Y_i) = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i + \beta_4 M_i$
2. $logit(Y_i) = \beta_0 + \beta_1 (A * M)_i + \beta_2 B_i + \beta_3 C_i$
3. $logit(Y_i) = \beta_0 + \beta_1 A_i + \beta_2 (B * M)_i + \beta_3 C_i$
4. $logit(Y_i) = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 (C * M)_i$

where Y is the odds ratio of a positive outcome for individual i, the β terms are vectors of varying sizes relating to each factor, and $*$ represents terms being interacted.

171. The AIC statistic is output for each model. If the AIC for either model 2, 3 or 4 is smaller than that of model 1 (which has no interaction terms), then that model is inspected further. If an interaction term between factor M and another factor shows significance, then we consider this factor against the primary factors. Due to the limits of the benchmarking process on the number of factors we can keep, the parameter estimates would need to indicate that the interaction was particularly important for factor M to be retained over and above a primary factor.

172. We think that this approach strikes a balance between thoroughness and practicality. Since we have so many potential combinations of interactions to consider, across many different measures, we need an approach which minimises manual processing, and removes subjectivity as far as possible.

173. We are not considering three-way or higher order interactions. If a secondary factor has not shown significant effects either on its own or in interaction with any of the primary factors, then we deem it unlikely that it will be significant at higher levels of interaction. Additionally, it is questionable if a factor can reasonably be said to be correlated with the outcome if it is only correlated in combination with two other factors.

174. In practice, there were a few cases where factors showed no significant estimated differences on their own but significantly improved the model fit when interacted with another factor. However, in no case did the parameter estimates indicate that these relationships were enough to prioritise retaining the secondary factor over a primary factor with larger significant estimated differences.

[46] Akaike Information Criterion; a measure of model fit adjusted for number of independent variables.