Annex B: Statistical modelling

Methodology overview

1. This annex outlines the methodology used for the statistical modelling of the attainment of first and upper second class degrees combined, and of first class degrees alone.

Graduate population

2. The graduate population in this analysis comprises UK-domiciled first degree graduates who studied full-time, were registered at higher education providers in England and graduated in the academic years from 2010-11 to 2020-21 with a classified honours degree.

3. This population (Graduate population A) is summarised in Table 3 of Annex A and can be rebuilt using the following fields described in the OfS publication 'Technical algorithms for institutional performance measures':[1]

   - DFAPAPPEXCL = 0
   - IPDOQUALPOP = 1
   - IPCOUNTRY = 'E'
   - IPEMPLEVEL in ('DEG', 'PUGD')
   - IPEMPMODE = 'FT'
   - IPBASEYEAR in (2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020).

4. In our analysis this population was further limited to those who qualified from English providers awarding at least 10 classified honours degrees in each of the academic years considered. This population (Graduate population B) is summarised in Table 4 of Annex A.

Method to determine 'unexplained' attainment

5. Mixed-effects logistic regression modelling was employed to investigate whether or not the observed changes in graduate attainment over time at the sector and provider levels can be explained by changes in the makeup of the graduate population in terms of the explanatory variables included in the modelling.

6. The modelling used to investigate degree attainment changes with time at the sector level includes explanatory variables relating to the provider at which the graduate was registered, graduation year and various key graduate characteristics. The effects of the following were included as explanatory variables in the model:

   - the provider at which the graduate was registered

[1] See 'Technical algorithms for institutional performance measures: Core algorithms', available under 'OfS core algorithms' at www.officeforstudents.org.uk/data-and-analysis/access-and-participation-data-dashboard/guide-to-the-access-and-participation-data-resources/.

- year of graduation
- subject of study
- qualifications on entry into higher education
- age
- declared disability status
- ethnicity
- sex
- tracking underrepresentation by area based on Middle Super Output Areas (TUNDRA MSOA) quintile[2].

7. We have created an update to the model we published previously (the 'original' model).[3] The structure of this updated model is more complex than used in previous reports, as we have tried to account for additional changes in the time series related to attainment during the coronavirus pandemic. We now interact subject of study, entry qualifications, age, disability status, ethnicity, sex and TUNDRA MSOA quintile with year of graduation to account for changes in awarding, especially since the start of the coronavirus pandemic (the 'added interactions' model). These interactions can be seen in Equation B1.

8. Our model allows us to predict the proportion of graduates awarded a first or an upper second class degree, or a first class degree, accounting for the effects of the explanatory variables.

9. To investigate and isolate the effect of graduation year on degree attainment, the following methodology was applied:

    a. The optimised models provide the probability of an individual with given characteristics attaining a first or upper second class degree, or a first class degree.

    b. The predicted probability for a given group of individuals (e.g. white female students graduating in 2011-12) may then be determined by taking the mean of the predicted probabilities of the individuals in that group.

    c. To investigate the effect of graduation year on degree classification attainment, the model is applied to the entire reported graduate population, but with the academic year of graduation for all graduates in the population changed to 2010-11.

[2] See www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/about-tundra/.

[3] Details of the 'original model' can be found in the previous version of this report, 'Analysis of degree classifications over time: Changes in graduate attainment from 2010-11 to 2018-19' (OfS 2020.52, available at www.officeforstudents.org.uk/publications/analysis-of-degree-classifications-over-time-changes-in-graduate-attainment-from-2010-11-to-2018-19/).

d. The observed value for the proportion of graduates attaining a first or upper second class degree, or a first class degree, in each academic year is then compared with the model's predicted value for the same graduates, had they graduated in 2010-11.

e. Any differences between the predicted and observed values is said to be 'unexplained', and a result of unobserved effects between academic years that have not been accounted for and have not been included as explanatory variables in the model. It is not possible to determine from this analysis what these additional unobserved factors are.

10. In summary, we estimate the 'unexplained' difference in the proportion of graduates attaining a first or upper second class degree, or a first class degree, had they graduated in 2010-11, compared with the actual year of their graduation.

Hypothetically closed attainment gaps within additional contextual variable groups

11. Additionally, we have applied the same method presented in paragraphs 5 to 8 of this annex but have further assigned all graduates the values for the additional contextual variables (sex, ethnicity, disability status and TUNDRA MSOA quintile) associated with the groups that have the highest attainment when graduating in 2020-21.

12. For firsts and for first and upper second class degrees combined, at a sector level, the highest attaining graduates in 2020-21 for the additional contextual variable groups were white, non-disabled female students from TUNDRA MSOA quintile 5 areas.

13. The predicted attainment of the graduate population in 2020-21 may be considered a hypothetical upper estimate of the expected sector attainment, representing a hypothetical sector where attainment gaps between groups within the additional contextual variable groups do not exist.

Model details

14. Here we detail the model used to describe the attainment of first and upper second class degrees combined, and of first class degrees alone.

15. Mixed-effects logistic regression has been used to model the probability of graduate i attaining a first or an upper second class degree, or a first class degree, from provider j, accounting for the effect of the explanatory variables outlined in paragraph 6 of this annex.

16. Following the changes in attainment of particular groups of students during the coronavirus pandemic (see paragraphs 18 to 21 in the 'Context and background' section in the main report) we have added 'year' interaction terms to the standard model previously used for this report. This is a more complex model, with the addition of interaction terms, to better capture year-on-year sector changes in attainment for particular student characteristics.

17. The specification of the statistical model used in the analysis is displayed in Equation B1.

Equation B1: Mixed-effects logistic regression model for graduate degree attainment

$$\boldsymbol{first\ or\ upper\ second\ class\ OR\ first\ class} \sim \boldsymbol{Binomial}(\boldsymbol{n_{ij}}, \boldsymbol{\pi_{ij}})$$

$$\begin{aligned}\boldsymbol{logit}(\boldsymbol{\pi_{ij}}) = {} & \boldsymbol{\beta_{0j}} + \boldsymbol{u_{0j}} + \sum_{\boldsymbol{Y=1}}^{\boldsymbol{11}} (\boldsymbol{\beta_Y} + \boldsymbol{u_{Yj}})\boldsymbol{X_{Yij}} + \sum_{\boldsymbol{Sbj=1}}^{\boldsymbol{21}} \boldsymbol{\beta_{Sbj}}\boldsymbol{X_{Sbjij}} + \sum_{\boldsymbol{Q=1}}^{\boldsymbol{23}} \boldsymbol{\beta_Q}\boldsymbol{X_{Qij}} + \sum_{\boldsymbol{A=1}}^{\boldsymbol{2}} \boldsymbol{\beta_A}\boldsymbol{X_{Aij}} \\ & + \sum_{\boldsymbol{D=1}}^{\boldsymbol{2}} \boldsymbol{\beta_D}\boldsymbol{X_{Dij}} + \sum_{\boldsymbol{E=1}}^{\boldsymbol{6}} \boldsymbol{\beta_E}\boldsymbol{X_{Eij}} + \sum_{\boldsymbol{Sex=1}}^{\boldsymbol{3}} \boldsymbol{\beta_{Sex}}\boldsymbol{X_{Sexij}} + \sum_{\boldsymbol{T=1}}^{\boldsymbol{6}} \boldsymbol{\beta_T}\boldsymbol{X_{Tij}} \\ & + \sum_{\boldsymbol{(Q*A)=1}}^{\boldsymbol{46}} \boldsymbol{\beta_{(Q*A)}}\boldsymbol{X_{(Q*A)ij}} + \sum_{\boldsymbol{(Y*Sbj)=1}}^{\boldsymbol{231}} \boldsymbol{\beta_{(Y*Sbj)}}\boldsymbol{X_{(Y*Sbj)ij}} + \sum_{\boldsymbol{(Y*Q)=1}}^{\boldsymbol{253}} \boldsymbol{\beta_{(Y*Q)}}\boldsymbol{X_{(Y*Q)ij}} \\ & + \sum_{\boldsymbol{(Y*A)=1}}^{\boldsymbol{22}} \boldsymbol{\beta_{(Y*A)}}\boldsymbol{X_{(Y*A)ij}} + \sum_{\boldsymbol{(Y*D)=1}}^{\boldsymbol{22}} \boldsymbol{\beta_{(Y*D)}}\boldsymbol{X_{(Y*D)ij}} + \sum_{\boldsymbol{(Y*E)=1}}^{\boldsymbol{66}} \boldsymbol{\beta_{(Y*E)}}\boldsymbol{X_{(Y*E)ij}} \\ & + \sum_{\boldsymbol{(Y*Sex)=1}}^{\boldsymbol{33}} \boldsymbol{\beta_{(Y*Sex)}}\boldsymbol{X_{(Y*Sex)ij}} + \sum_{\boldsymbol{(Y*T)=1}}^{\boldsymbol{66}} \boldsymbol{\beta_{(Y*T)}}\boldsymbol{X_{(Y*T)ij}}\end{aligned}$$

Note: A number of the terms included in this equation are zero, as they refer to a reference category (see Table B1 for detail of reference categories).

18. In Equation B1 the βs represent the fixed effect coefficients that are common to individuals across all providers (the sector) and years,[4] Xs (0 or 1) represent whether or not an individual has the characteristics (Y = academic year of graduation, Sbj = subject of study, Q = entry qualifications, A = age, D = declared disability status, E = ethnicity, Sex = sex, T = TUNDRA MSOA quintile, $Q * A$ = interaction between entry qualifications and age, $Y * Sbj$ = interaction between academic year of graduation and subject of study, $Y * Q$ = interaction between academic year of graduation and entry qualifications, $Y * A$ = interaction between academic year of graduation and age, $Y * D$ = interaction between academic year of graduation and declared disability status, $Y * E$ = interaction between academic year of graduation and ethnicity, $Y * Sex$ = interaction between academic year of graduation and sex, $Y * T$ = interaction between academic year of graduation and TUNDRA MSOA quintile), u_{0j} is the random intercept for provider j and u_{Yj} represents the random coefficient for provider j in academic year Y with

$$u_{0j} \sim \mathrm{N}(0, \sigma_{u_0}^2)$$

$$u_{\mathrm{Y}j} \sim \mathrm{N}(0, \sigma_{u_Y}^2).$$

19. A full summary of the variables used in the model, and the categories within those variables, is given in Table B1. Table B1 does not include the various interaction terms featured in the model (581 non-zero estimates). Each year is interacted with the student characteristics independently, resulting in the large number of interaction terms.

[4] The summation term for academic year of graduation includes the reference year of 2010-11, as each provider has a random coefficient for all years but the fixed effect 2010-11 coefficient $\beta_1 = 0$ (reference categories for other explanatory variables are omitted from the model structure).

20. Fixed effect coefficient estimates, standard errors and p-values for models can be found in Table 6 of Annex A. Estimates for the interactions terms are not included in this table. They are available on request.

Table B1: Variables used in the graduate degree attainment modelling (all categorical)

Model variable name	Description
Academic year (Y)	Academic year of graduation: 2010-11 (ref) 2011-12 2012-13 2013-14 2014-15 2015-16 2016-17 2017-18 2018-19 2019-20 2020-21
Subject of study (Sbj)	Subject studied: Agriculture, food and related studies Architecture, building and planning Biological and sport sciences Business and management (ref) Combined and general studies Computing Design, and creative and performing arts Education and teaching Engineering and technology Geography, earth and environmental studies Historical, philosophical and religious studies Language and area studies Law Mathematical sciences Media, journalism and communications Medicine and dentistry Physical sciences Psychology Social sciences Subjects allied to medicine Veterinary sciences
Entry qualifications (Q)	Entry qualifications of the graduate: A-level: AAA and above (ref) A-level: AAB A-level: ABB

Model variable name	Description
	A-level: BBB A-level: BBC A-level: BCC A-level: CCC A-level: CCD A-level: CCD A-level: CDD A-level: DDD A-level: Below DDD BTEC: DDD and above BTEC: DDM BTEC: DMM BTEC: MMM and below 2 A-levels and 1 BTEC 1 A-levels and 2 BTEC International Baccalaureate Other Level 3 No Level 3 Equivalent
Age (A)	Age on entry Under 21 (Young) (ref) Over 21 (Mature)
Disability (D)	Declared disability status of graduate Disability No disability (ref)
Sex (Sex)	Sex of graduate: Female (ref) Male Other
Ethnicity (E)	Ethnicity of graduate: Asian Black Mixed Other White (ref) Unknown
TUNDRA MSOA quintile (T)	Young participation quintile of graduate: Quintile 1 Quintile 2 Quintile 3 (ref for firsts only model) Quintile 4 (ref for firsts or upper second model) Quintile 5 Unknown

Note: Those categories marked with '(ref)' are the reference categories for each categorical or dummy variable and are not formally included in the model structure (they are equal to 0).

21. In changing the composition of the underlying statistical model, estimates for unexplained attainment are likely to have changed from previously published figures. We have published new provider-level estimates in Annex A Tables 1 and 2. Figures B1 and B2 show the sector-level differences in attainment estimates between the model used in previous reports (the 'original' model) and the model we present in this report (the 'added interactions' model) for first and upper second class and first class degree attainment.

Figure B1: Sector-level first and upper second class attainment observed and modelled from the 'original' and 'added interactions' models between 2010-11 and 2020-21

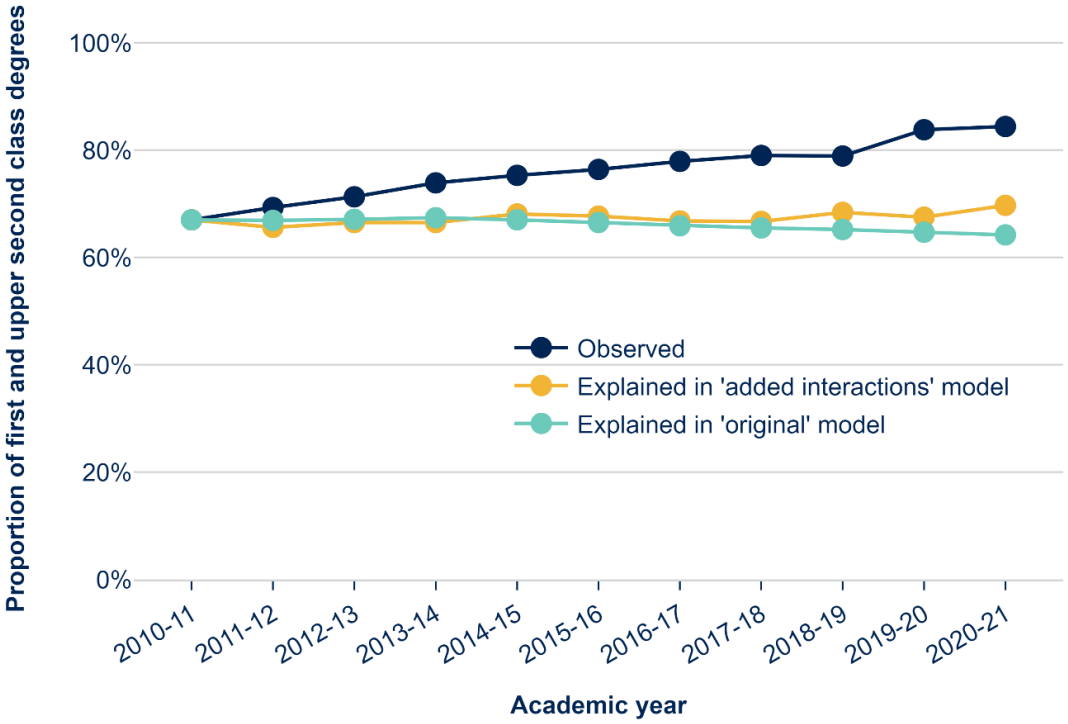
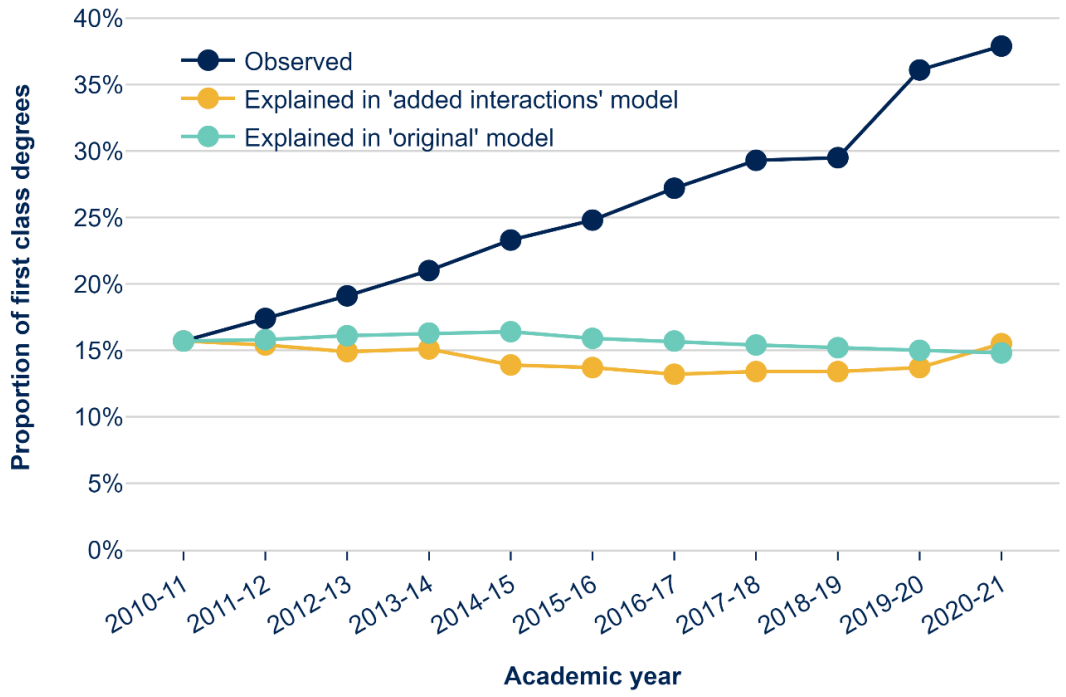


Figure B2: Sector-level first class attainment observed and modelled from the 'original' and 'added interactions' models between 2010-11 and 2020-21



22. Figures B1 and B2 show that the unexplained attainment estimates (the difference between the observed and explained attainment) using the 'added interactions' model are slightly different from the estimates from the 'original' model. The 'added interactions' model is more complex and is better at capturing changes over time.

23. The model outputs presented in Annex A for first and upper second class (Table 1) and first class only (Table 2) are from the 'added interactions' model.

24. Estimates of the variance components and their standard errors for the random intercepts and random year coefficients are shown for first and upper second class and first class degrees in Table B2.

Table B2: Variance component estimates for the models for first or upper second class and first class only degree attainment

	Random effect	First or upper second class model estimate	First or upper second class model standard error	First class model estimate	First class model standard error
Intercept	$\sigma^2_{u_0}$	0.123	0.016	0.167	0.021
Year	$\sigma^2_{u_Y}$	0.034	0.002	0.038	0.002

25. Model fit statistics for first and upper second class and first class degrees can be found in Table B3.

Table B3: Model fit statistics for the models for first or upper second class and first class only degree attainment

Statistic	First or upper second class model	First class model
-2logLiklihood	2,606,846	2,792,084
Akaike information criterion	2,608,148	2,793,386