# Higher Education Learning Gain Analysis (HELGA)

## Can administrative data be used to measure learning gain?

**Reference** OfS 2019.07

**Enquiries to** annalise.ruck@officeforstudents.org.uk

**Publication date** 12 July 2019

# Table of Contents

# Summary

1. This publication sets out the work of the Higher Education Learning Gain Analysis (HELGA) strand of the learning gain programme led by the Office for Students (OfS). HELGA aimed to assess whether or not it is possible to use administrative data to measure learning gain. In doing so, institutional value-added measures have been created using two different statistical techniques: multilevel modelling and a pairwise comparison method.

2. Discussion around how value-added can be measured, which predictor variables and output variables to use and whether contextual variables should additionally be adjusted for are included in this report. It also sets out work that has been done to adjust UCAS tariff points to account for some A-level subjects being more difficult than others.

3. HELGA has not succeeded in finding a single measure of learning gain that could be used across the sector based on administrative data. However, this body of work explains the various avenues that have been explored and the results of these analyses. Therefore, it provides points for further discussion as the sector continues to look for ways that learning gain can be measured.

# 1. What is Higher Education Learning Gain Analysis (HELGA)?

4. The Office for Students' (OfS) learning gain programme was created to explore the many ways in which the sector can measure what is gained by students from their higher education experience. This is complex and multi-faceted as there are many ways in which students progress during their time at university. Elements include improvements to knowledge, skills, work-readiness and personal development.

5. The OfS learning gain programme was commenced by the Higher Education Funding Council for England (HEFCE) before being inherited by the OfS in April 2017. It is framed by the report from RAND Europe, commissioned by HEFCE. The report defined learning gain as: 'the difference between the skills, competencies, content knowledge and personal development demonstrated by students at two points in time[1].' This definition immediately highlights the complexity of measuring learning gain in highlighting the many different elements that must be considered.

6. The learning gain programme has three strands: pilot projects, the National Mixed Methodology Learning Gain (NMMLG) project, and the Higher Education Learning Gain Analysis (HELGA). The pilot projects and NMMLG project sought to develop and test instruments that measure various aspects of these skills and knowledge[2]. To complement that work, HELGA examines existing administrative data on students' experience to evaluate whether it can be used to deepen the understanding of learning gain.

7. RAND's definition of learning gain relies on measuring skills on the same scale multiple times so RAND distinguishes between learning gain and value-added: 'Learning gain is measured based on the difference between two measures of actual student performance, while value added is based on the comparison between performance predicted at the outset of studies and actual performance results.'

8. Since there is no such single scale in the higher education system, HELGA necessarily focuses on developing a measure of value-added as a proxy measure for learning gain, which will allow for comparison between institutions. Any proxy measure created using administrative data is likely to have less validity than a measure based on a bespoke data collection. However, collection of bespoke data would require a lot of resource and so if a valid measure could be generated from existing data, this is likely to be preferable, even if it is less accurate.

9. To support the development of this work, an expert group, made up of specialists in the area of measurement of educational progress, was established. The group has advised on various aspects of the methodology, including discussing the most appropriate outcomes measures, data sources and technical aspects of the modelling methodologies. Details of the group

[1] McGrath, C.H., Guerin, B., Harte, E., Frearson, M. and Manville, C., 2015. Learning gain in higher education. *Santa Monica, CA: RAND Corporation*. Available from: www.officeforstudents.org.uk/media/11b42adc-534c-481e-91e9-aa87fbddff62/learning-gain-rand-report.pdf [PDF]

[2] For further details of these strands see: www.officeforstudents.org.uk/advice-and-guidance/teaching/learning-gain/

membership can be found on the OfS website[3]. We would like to take this opportunity to thank the members of the expert group for their advice and guidance during the course of this project.

10. This report describes the steps taken to explore the different options for creating a value-added measure. It starts with discussion around how students' outcomes can be measured and what variables might be used to predict those outcomes. This is followed by sections on statistical modelling for adjusting UCAS tariff points to account for some A-level subjects being more difficult than others, the statistical modelling used for creating a value-added measure, and a brief evaluation of the HELGA project.

[3] Available at: www.officeforstudents.org.uk/advice-and-guidance/teaching/learning-gain/

# 2. How HELGA measures value-added

11. Estimating value-added from administrative data involves predicting an outcome and then calculating the difference between the observed and predicted outcome. There are many approaches to predictive modelling and many ways to estimate the value-added model. Common to all approaches are two conceptual problems: which variables should be used to predict the outcome? And what outcomes should be measured?

12. Decisions need to be made about which students will be included in the analysis, decisions which are likely to be influenced by the selected predictor and outcome measures. This is because data on all possible predictor or outcome measures is not available for all students.

13. Consideration also needs to be given to which student contextual variables should be included. Contextual variables, such as students' characteristics, allow the performance of the institution to be isolated. This may be particularly important since it is well established that outcomes differ across different student groups[4].

## 2.1. Data sources

14. The decisions to be made around which outcome measure and predictor of outcomes to use will be largely based on the availability of data. The main source of administrative data held by the OfS is the Higher Education Statistics Agency (HESA) individualised student record[5]. This collects information about the attributes of each individual higher education student registered at a higher education provider in the UK in a given year, as well as details of the study undertaken and any qualifications achieved.

15. The OfS also holds data from the Destinations of Leavers from Higher Education (DLHE) survey and the Longitudinal Destinations of Leavers from Higher Education (LDLHE) survey[6]. These survey graduates at six months and 40 months after graduation respectively to identify the activities undertaken since graduation. Activities include full-time or part-time employment, further study or being unavailable for employment. These surveys also give details of the activity being undertaken, such as the type of work or the level of study.

16. Additionally, the OfS holds all responses to the National Student Survey (NSS), elements of which can be considered for measuring outcomes.

## 2.2. Choosing a predictor of outcome

17. In the English education system there is no standard measure of performance or attainment at the end of secondary or tertiary education. This makes choosing a measure that can predict performance in higher education difficult. Of course, the majority of students entering higher education will have some kind of qualification obtained before starting their course. However, using entry qualifications as the predictor of outcome accounts only for the starting point of a student's cognitive skills. Without access to data that might allow measurement, or inclusion, of non-cognitive skills there will be a necessary focus on cognitive skills.

[4] See www.officeforstudents.org.uk/data-and-analysis/differences-in-student-outcomes/
[5] See www.hesa.ac.uk/data-and-analysis/students
[6] See www.hesa.ac.uk/data-and-analysis/graduates

18. For students from England entering higher education aged 21 or under, the majority will do so with three or more A-levels[7]. A-levels are known to have a strong positive relationship with degree outcomes, and degree outcomes are known to be correlated with other outcomes, such as employment outcomes[8]. Therefore, A-levels are a good choice of predictor of outcomes and so they will be used in this analysis. Specifically, the UCAS tariff points associated with A-level grades will be used. Since not all students hold A-levels when entering higher education, this reduces the number of students in the analysis, the details of which can be found in Section 4.

19. Section 3 describes a statistical modelling technique that explores whether all A-level subjects are of the same level of difficulty and how tariff points can be adjusted to account for any discrepancies in difficulty to ensure that the 'starting point' for students isn't unfairly influenced by their subjects studied at A-level.

20. The expert group also suggested that GCSEs should be considered as the predictor of outcome. When choosing whether or not to offer a student a place at university, admissions teams will often consider their prior attainment. Up until recently, this would have included a combination of GCSE and AS-level grades. However, the reforms to A-levels from September 2015[9] have meant that universities will not have the same access to AS grades to assist with this decision making. Due to this, the Department for Education (DfE) has carried out analysis to assess which of these measures (GCSEs or AS-levels) are better predictors of later performance[10].

21. This research found that neither GCSE nor AS-level results alone were able to predict degree classification with great accuracy, but that GCSEs were the better of the two at making these predictions. However, research from the University of Cambridge[11] and from the University of Bristol[12] has found this might not be the case.

22. Carrying out this analysis using GCSEs as the predictor is outside the scope of this project. However, the models used to create the value-added measure would allow for GCSEs to be considered rather than, or as well as, A-levels.

## 2.3. Choosing an outcome measure

23. Outcome measures available in the administrative data include degree classification, employment status six months after graduation and salary six months after graduation. For some cohorts, there is also a record of employment status and salary 40 months after

[7] For example, for the students included in this analysis (see Section 4 for details) 143,855 of the 212,835 (67.6 per cent) held three or more A-levels when they started their higher education course.
[8] See www.officeforstudents.org.uk/data-and-analysis/differences-in-student-outcomes/degree-outcomes-overview/
[9] See www.gov.uk/government/publications/get-the-facts-gcse-and-a-level-reform/get-the-facts-as-and-a-level-reform
[10] See https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/200903/GCSE_and_AS_level_Analysis_3_1.pdf [PDF]
[11] Partington, R., Carroll, D. and Chetwynd, P., 2011. Predictive effectiveness of metrics in admission to the University of Cambridge. *University of Cambridge.* Available at: https://www.cao.cam.ac.uk/sites/www.cao.cam.ac.uk/files/ar_predictive_effectiveness_of_metrics_in_admission.pdf [PDF]
[12] Available at: http://blogs.lse.ac.uk/impactofsocialsciences/2014/07/30/replicating-government-commissioned-research/

graduation. Use of metadata from the NSS survey to measure non-cognitive abilities has also been considered.

24. As mentioned above, employment outcomes at six months are obtained from the DLHE survey. Use of this data would mean that not all students in our chosen student group could be included in the analysis. This is because not all graduates respond to the survey – for UK-domiciled and EU-domiciled graduates, the response rate for the most recent graduates (2016-17) was 77 per cent[13]. Using employment outcome – that is, whether or not a student was in highly skilled employment, or in any kind of employment – would reduce the number of graduates in the analysis from 138,325 to 84,580. This would be reduced even further if employment outcomes after 40 months, from the LDLHE, were used. Whilst it might be possible to consider imputing the missing data, no analysis has been carried out as to whether or not this data is missing at random or the impact of imputation as part of this project. Additionally, a graduate's employment outcome could be affected by a number of factors outside the institution's influence.

25. New research suggests that non-cognitive performance might be objectively measurable using the information in the metadata for survey responses[14]. For example, skipping questions in a survey may correlate with non-cognitive performance. Metadata for the NSS has been examined to test whether the appropriate data is available to develop such a measure of non-cognitive performance.

26. Three types of behaviour have been identified that can be associated with conscientiousness (or its lack). These were: skipping of questions or quitting the survey early; yea-saying[15] and responding disproportionately quickly (or slowly). However, it was not possible to identify these behaviours because of the validation rules in place on the online version of the survey, which was the method used by 80 per cent of respondents. The online survey does not allow students to skip questions, so the first behaviour was not measurable. Additionally, the metadata on response times was not available, meaning this approach is not viable.

27. Given the above, degree classification is the best available outcome measure for this analysis. However, using degree classifications is not without its problems. Firstly, degree classifications are not available for all students. In particular, graduates from courses in veterinary sciences and medicine do not receive a degree classification and are recorded as having an 'unclassified' degree. Additionally, there is no standardised curriculum across institutions and so it cannot be certain whether a first-class degree from institution A is the same as a first-class degree from Institution B[16]. Despite this, the expert group were in agreement that degree classification was the most appropriate choice from the available outcome measures to use for testing these methodologies for a measure of value-added.

[13] See https://www.hesa.ac.uk/news/28-06-2018/sfr250-higher-education-leaver-statistics

[14] Jackson, C. Kirabo. 'What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes'. Working Paper. National Bureau of Economic Research, May 2016. Available from: http://www.nber.org/papers/w22226

[15] 'Yea-saying' is where a respondent gives positive answers to all questions, even when they are uncertain of their true response.

[16] The QAA sets out subject benchmark statements, which show what graduates can expect to know, do and understand at the end of their studies (https://www.qaa.ac.uk/quality-code/subject-benchmark-statements#), but it is still the case that curricula vary greatly while sitting within these statements.

## 2.4. Choosing who to include in the analysis

28. Since the aim of HELGA is to assess whether or not it is possible to use administrative data to create a measure of value-added, we are not attempting to create a measure for all students studying at all levels at this stage. Instead, this project aims to create a framework for measuring value-added and to assess the validity of that measure. Should the measure be deemed to be appropriate and effective, then this framework could be developed further in order to include a wider diversity of students.

29. Therefore, this analysis focuses on entrants to full-time first degree courses at publicly funded higher education institutions who were UK-domiciled at the start of their course. We have further restricted this group to only those who were 18 or 19 at the start of this course. This differs from the usual definition of 'young' learners to ensure that as many of these entrants as possible will have achieved their Level 3 qualifications in the same academic year.

30. Given that an outcome measure will be required for the value-added measure, it is necessary to track entrants through their courses until they complete it, or are no longer active on the course. Since not all courses are the same length, the decision has been taken to allow six years for students to complete their course, which should allow for most of the longest courses. The year of entry used is 2009-10, giving students until the year 2014-15 to complete their course.

## 2.5. Contextual variables: should they be included?

31. The question of whether or not to include contextual variables in a value-added measure is a complex one. The idea behind including contextual variables is that, since it is well established that nationally some groups of students perform worse than others, even after accounting for their prior attainment, institutions with a higher proportion of students from these groups should not be penalised for performing less well than institutions with fewer of those students. This is particularly important since it is often the case that the student groups who perform less well are from disadvantaged backgrounds. Since we want to encourage students from disadvantaged backgrounds to attend university, we do not want to discourage institutions from admitting these students by making comparisons of progress that do not account for the make-up of the student body. Parallel arguments have long been made for secondary school value-added measures[17].

32. However, there is concern that including contextual variables in value-added measures might allow institutions to set a lower bar for students who typically perform less well in higher education. Since there are no standard value-added measures for higher education, it is difficult to understand if there is an established approach to the inclusion or exclusion of contextual variables. However, we can consider what has happened in the measurement of value-added in secondary education.

[17] Leckie, G. and Goldstein, H. (2019) The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England. British Educational Research Journal. DOI: 10.1002/berj.3511. Available from: https://onlinelibrary.wiley.com/doi/10.1002/berj.3511

33. School progress tables, published annually by the DfE, include a measure of school progress[18]. Since 1992, this measure has taken four forms: value-added, contextual value-added, expected progress and Progress 8. The move from value-added to contextual value-added followed a report from the National Audit Office in which they stated that data on the 'academic achievements of the pupils in earlier stages of their education and on aspects of their economic, social and cultural backgrounds' should be used when measuring school progress and that doing so would provide 'a more robust and objective assessment of the relative performance of schools'[19].

34. However, in 2010 the contextual value-added measure was scrapped. Part of the reason for this was that the government felt that it was wrong to expect different levels of progress from different groups of students. They said that 'It is morally wrong to have an attainment measure which entrenches low aspirations for children because of their background'[20]. In their critique of the evolution of school league tables[21], Leckie and Goldstein argue that this view is a minority one in the academic literature. The more common argument is that, given that nationally some pupil groups make less progress than others, this must be adjusted for if fair comparisons are to be made between schools.

35. In this report, where possible, results will be presented both with and without contextual variables. No view is expressed as to which is correct as we believe this requires further research and discussion beyond the scope of this project.

## 2.6. Statistical techniques

36. The HELGA project has explored two techniques that can be applied to most predictors and outcome measures to estimate value-added. When applied to measures of students' attainment, they yield a comparison of the value-added across institutions. These techniques and their associated insights are:

    - Multilevel modelling to partition the variance in value-added between the institution and the individual. This has the advantage of being applicable to nearly all students but can only account for characteristics observed in the administrative data, such as gender and ethnicity (the administrative data do not observe other characteristics which may have an effect – such as students' personal preferences of elements of study such as the location of the institution or the availability of extra-curricular activities)[22].

[18] Available at: https://www.gov.uk/government/statistics/secondary-school-performance-tables-in-england-2018-provisional

[19] See https://www.nao.org.uk/report/making-a-difference-performance-of-maintained-secondary-schools-in-england/

[20] DfE, U.K., 2010. The importance of teaching: The schools white paper. *White Papers), CM*, *7980*. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/175429/CM-7980.pdf [PDF]

[21] Leckie, G. and Goldstein, H., 2017. The evolution of school league tables in England 1992–2016:'Contextual value-added', 'expected progress' and 'progress 8'. *British Educational Research Journal*, *43*(2), pp.193-212. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/berj.3264

[22] For examples of the use of similar techniques see Broeke S, and Nicholls T, 2006, 'Ethnicity and degree attainment' available from: http://dera.ioe.ac.uk/6846/ HEFCE (2015/21), 'Differences in degree outcomes: The effect of subject and student characteristics' available from: http://www.hefce.ac.uk/pubs/year/2015/201521/

- A quasi-experimental, paired comparisons technique[23] that exploits the similarity between the students who were equally likely to be accepted or not accepted by an institution and those who are on the verge of admission but then enrol elsewhere. This allows many unobserved characteristics to be accounted for, such as preference of a location.

37. The second methodology requires use of UCAS admissions data. This restricts the coverage of the research because, whilst the vast majority of higher education providers use UCAS to recruit their students, not all of them do. In particular, some conservatoires use a separate system so a number of creative arts students will be omitted from the analysis. Additionally, some medical schools are run jointly across multiple providers meaning that it is not possible to attribute the learning gain of these students to one provider in particular. Details of the number of students included in this methodology can be found in Section 4.

[23] See Hoxby, C, 2015, 'Computing the Value-Added of American Postsecondary Institutions'. Available from: https://www.irs.gov/pub/irs-soi/15rpcompvalueaddpostsecondary.pdf [PDF]

# 3. Adjusting A-level tariff points

38. Analysis from Ofqual[24] has found that not all A-levels are of the same level of difficulty. For this analysis we wanted to be able to account for these discrepancies in difficulty. The calculation of value-added relies on the predictor variables' ability to predict the outcomes. Ensuring that these discrepancies in difficulty are accounted for should improve the accuracy of the predictive power of tariff points. To do this, we have developed a methodology based on concepts from Rasch modelling used in the Ofqual report.

39. This modelling only considers those with three or more known A-level grades and who were 18 at the start of their undergraduate degree. In order to ensure that there are sufficient numbers of students to run the analysis, four years of data have been included: from 2008-09 to 2011-12. The age restriction and controlling for academic year in the statistical model ensures that we are comparing grades for students who sat their A-levels in the same academic year. Additionally, A-level subjects were removed if there were fewer than 500 students holding A-levels in that subject.

40. The first method that was explored was a multilevel model using fixed effect dummy variables for academic subjects and a random effect for the individual student. The random effect seeks to represent the underlying ability of the individual.

41. The model used was:

$$Ucas\ tariff\ score_{ij} = \beta_0 + \sum_{j=1}^{s-1} \beta_j subject_{ij} + \sum_{l=s}^{s+(y-1)} \beta_l year_{il} + u_i + \epsilon_{ij}$$

Where: $i = individual$
$j = subject$
$s = number\ of\ subjects$
$l = year$
$y = number\ of\ years$
$u_i\ is\ the\ random\ effect\ for\ individual\ i.$

42. However, fitting random effects for every student entering higher education with A-level qualifications over a four-year period requires an extensive amount of computing power. Therefore, it has not been possible to obtain estimates from this model so a second methodology has been considered.

[24] For example, Ofqual, 2015. Comparability of different GCSE and a level subjects in England: An introduction: ISC (Working Paper No. 1). Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/606041/1-comparability-of-different-gcse-and-a-level-subjects-in-england-an-introduction.pdf [PDF], Ofqual 2018. Inter-subject comparability in A level sciences and modern foreign languages. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/757841/ISC_Decision_Document_20.11.18.pdf [PDF] and Ofqual, 2015, 'Inter-Subject Comparability of Exam Standards in GCSE and A-Level: ISC Working Paper 3'. Coventry, The Office of Qualifications and Examinations Regulation. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/486936/3-inter-subject-comparability-of-exam-standards-in-gcse-and-a-level.pdf [PDF]

43. In this second methodology, rather than fitting a random effect for each individual in order to account for their underlying ability, standardised A-level points for each student is calculated. This is done by subtracting the mean number of tariff points for an individual from their tariff points for each of their A-level subjects. This approach essentially removes the individual effects from the model. A linear model is then used to estimate the standardised tariff points as follows:

$$Standardised\ tariff\ points_{ij} = \beta_0 + \sum_{j=1}^{s-1} \beta_j subject_{ij} + \sum_{l=s}^{s+(y-1)} \beta_l year_{il} + \epsilon$$

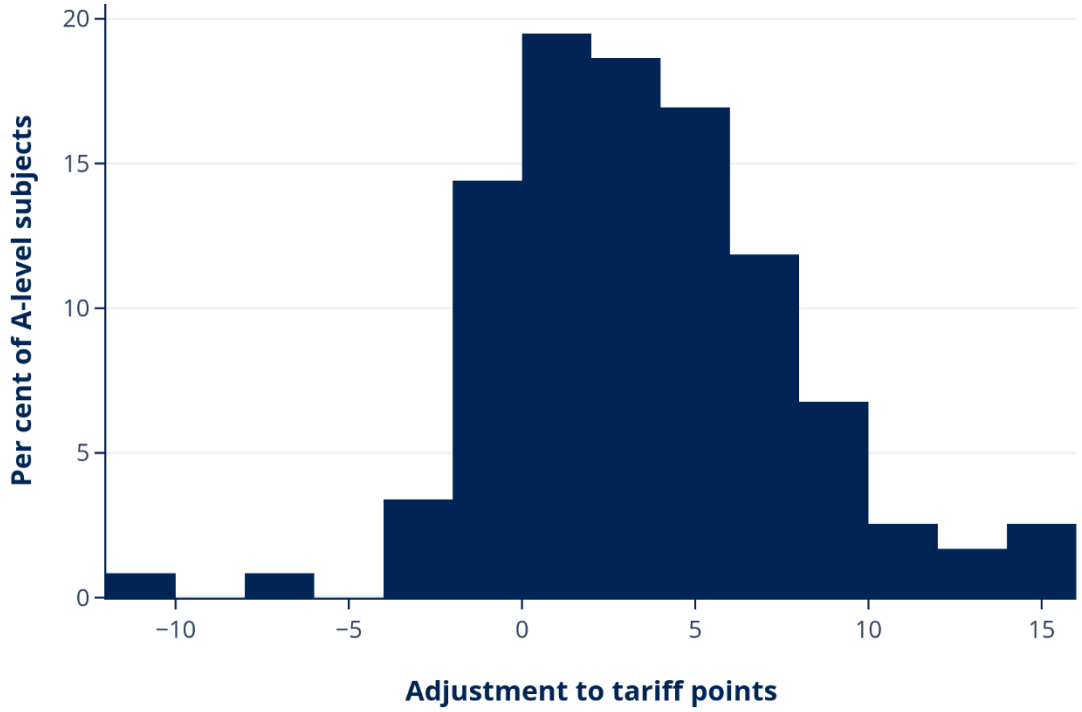
Where: $i = individual$
$j = subject$
$s = number\ of\ subjects$
$l = year$
$y = number\ of\ years.$

44. The model is weighted so that individuals with more than three A-levels do not have an undue influence on the model compared with those with three A-levels. It is assumed that no individual appears in more than one year.

45. History has been chosen as the reference category because it is a subject which a large number of students take and has average tariff points of 100 – the same as the average across all subjects. Estimates are produced for 113 A-level subjects, 103 of which were found to be statistically significant (at the α=0.05 level). For 16 subjects, the estimates were negative, meaning that points for those subjects will be increased (i.e., these subjects are found to be more difficult than history) and 87 were positive (i.e. less difficult).

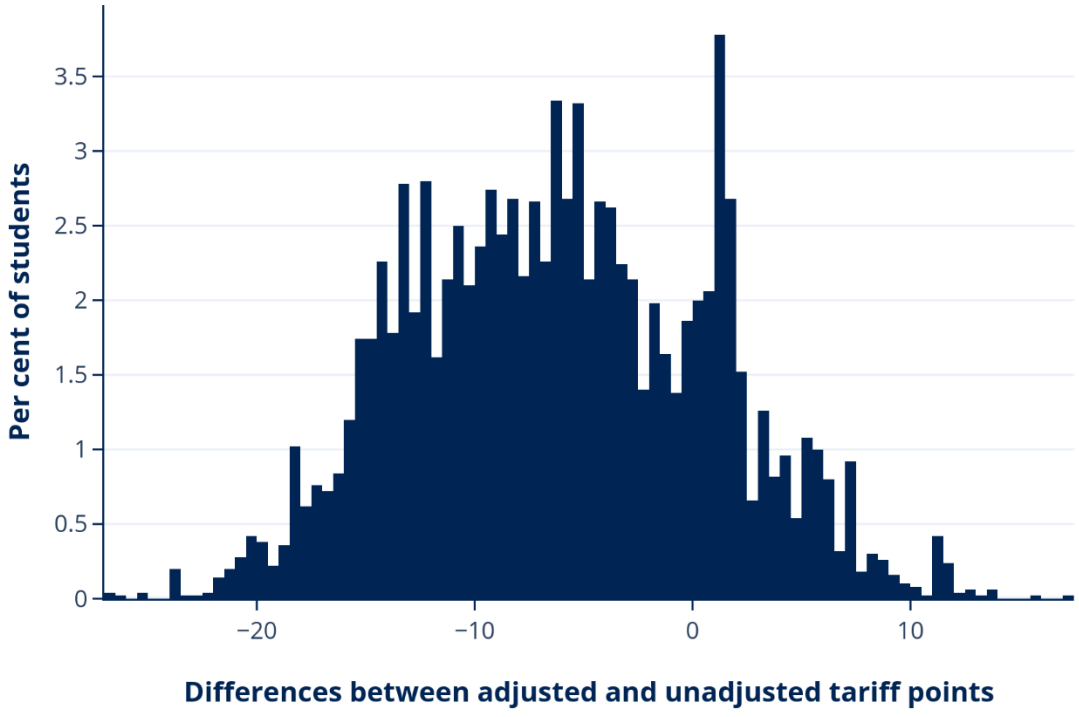
46. Figure 1 shows that, for the vast majority of subjects, the adjustment is less than 10 tariff points, equivalent to half a grade difference. There are no subjects for which the difference is greater than a whole grade (20 tariff points).

**Figure 1: Histogram of the adjustment in tariff points by subject**




47. Having calculated the estimates, the tariff points for each individual A-level a student holds have been adjusted (where estimates are available) and the total tariff points have been recalculated. The difference between the actual tariff points and the adjusted points are shown in Figure 2. From this, it can be seen that there are very few students for whom the points has been adjusted, up or down, by more than an A-level grade (less than one per cent).

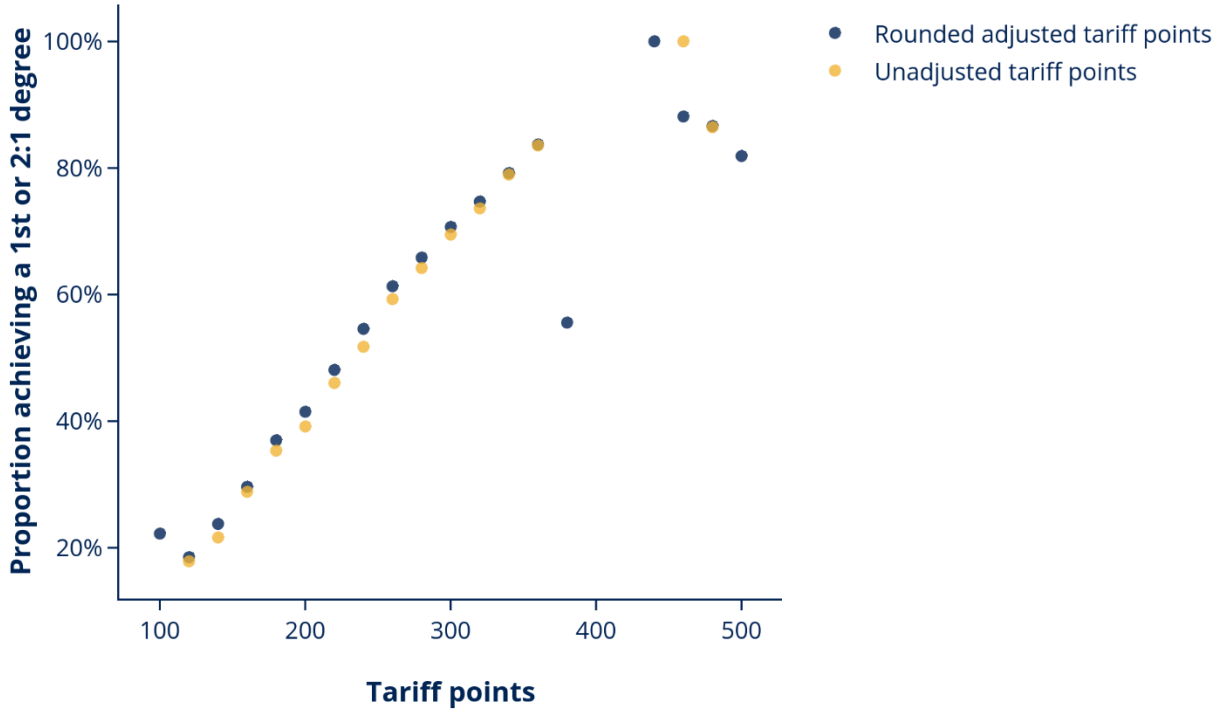
**Figure 2: Histogram of difference in adjusted and unadjusted tariff points by students**



48. However, for the purposes of this analysis, it is going to be necessary to look for students who have 'the same' number of tariff points. It is not possible to do this in large enough numbers with the adjusted tariff points and it is therefore necessary to group or round the adjusted points in some way. Since A-level tariff points are awarded in steps of 20 points, the decision has been taken to round the adjusted tariff points to the nearest 20. This results in around 18 per cent of students having tariff points that are adjusted by 20 points or more.

49. Since this rounding introduces a substantial difference in the number of students whose points change by 20 or more, all further analysis will be carried out separately for the unadjusted points and the rounded adjusted points to assess the difference this has on the modelling results. However, Figure 3 shows that the relationship between tariff points and degree classification differs very little once the rounded adjusted tariff points are used instead. This suggests that the model results are unlikely to differ when using either the adjusted or unadjusted tariff points.

**Figure 3: Percentage of students receiving a first or upper second class degree by rounded adjusted and unadjusted tariff points**




Note: The outlier at 380 points represents only 40 students and the outliers at 330 and 460 points represent fewer than 10 students.

50. Sensitivity analyses have been carried out to assess the impact of various aspects of the model including: grouping together four years of data, restricting the numbers of students in an A-level subject to 500 and restricting to only those who hold three or more A-levels. Details of these analyses can be found in Annex A.

# 4. Modelling value-added

51. This section describes in detail the previously introduced modelling techniques employed to calculate institutional value-added measures, namely, multilevel modelling and a pairwise comparison technique. The models include students who:

    - were entrants to first degree courses at English, publicly funded higher education institutions in 2009-10
    - were UK-domiciled and aged 18 or 19 at the start of their course
    - held three or more A-levels prior to entering higher education
    - had either completed their course or had left higher education without completing their course within six years of starting
    - completed their course and were awarded something other than an 'unclassified' degree.

    The pairwise comparison method student population is further restricted to those who applied to university through UCAS. Therefore, there are 138,325 students included in the multilevel modelling and 126,520 students included in the pairwise comparison modelling.

52. As discussed in Section 2.3, the outcome measure being used is degree outcome. Specifically, the proportion of students achieving a first or upper second class degree. That is, the proportion of all students who started a first degree course in 2009-10 and had either completed their course or left higher education without completing that course within six years. Students who were still active on their course after six years or who transferred to a different institution have been excluded from the analysis.

53. Students who receive an 'unclassified' degree tend to be studying courses such as medicine or veterinary sciences, where the typical degree classifications are not used, or are taking integrated masters courses, where classification can be awarded as Pass, Merit or Distinction. These students have been omitted from the analysis since there is no clear way to align their classifications with the standard undergraduate degree classifications.

## 4.1. The multilevel model

54. The first approach taken to estimating value-added is a two-level multilevel logistic regression modelling. This technique can account for various observed factors and allows for elements of the variation in value-added to be attributed at the institution level rather than just at the individual level. It is also possible to estimate departmental effects by modelling the effect of subject within institution. This leads to a three-level model. These models will enable the calculation of the predicted probability that a student will achieve a first or upper second class degree; using this information, alongside the actual proportion of students at a particular institution achieving these grades, a measure of the value-added attributed to the institution will be calculated.

55. The basic multilevel model being used is:

$$First\ or\ upper\ second_{ik} \sim Binomial(1, \pi_{ik})$$
$$logit(\pi_{ik}) = \beta_{0k} + \beta_1 adjusted\ tariff\ score_i + v_k$$

Where: $i = individual$
$k = insititution$
$v_k\ refers\ to\ unobserved\ institutional\ characteristics.$

56. Use of a multilevel model allows us to account for the fact that students will be studying in a department within a university. It is expected that each department and institution will impact on the students' probability of achieving a first or upper second class degree differently. In other words, we expect to find that the variance in degree outcome will differ between departments and institutions. Within this model, subjects have been grouped into 19 categories, which are considered as proxies for the departments in institutions, giving the following model:

$$First\ or\ upper\ second \sim Binomial(1, \pi_{ijk})$$
$$logit(\pi_{ijk}) = \beta_{0k} + \beta_1 adjusted\ tariff\ score + v_k + u_{jk}$$

Where: $i = individual$
$j = subject\ of\ study$
$k = insititution$
$v_k\ refers\ to$ unobserved $institutional\ characteristics$
$u_{jk}\ refers\ to$ unobserved $departmental\ characteristics.$

57. As discussed in Section 2.5, there is some debate around whether or not contextual variables should be included in these models. In order to assess the impact of including some student characteristics, these models will also be carried out with the following additional student-level variables: sex, disability, ethnicity and POLAR4. The Participation of Local Areas (POLAR) classification assigns small areas across the UK to one of five groups based on the proportion of the young population that participates in higher education[25].

58. In all, 10 different models have been considered:

- Null model with random intercepts for institution
- Fixed effect for tariff points, random intercepts for institution
- Fixed effect for adjusted tariff points, random intercepts for institution
- Fixed effects for tariff points, sex, disability, ethnicity and POLAR4, random intercepts for institution
- Fixed effects for adjusted tariff points, sex, disability, ethnicity and POLAR4, random intercepts for institution
- Null model with random intercepts for institution and subject
- Fixed effect for tariff points, random intercepts for institution and subject
- Fixed effect for adjusted tariff points, random intercepts for institution and subject

[25] For more information on POLAR, see www.officeforstudents.org.uk/data-and-analysis/polar-participation-oflocal-areas/

- Fixed effects for tariff points, sex, disability, ethnicity and POLAR4, random intercepts for institution and subject

- Fixed effects for adjusted tariff points, sex, disability, ethnicity and POLAR4, random intercepts for institution and subject.

59. To test the impact of the random effects, a model with no fixed or random effects has been carried out (a null model). The deviance of this model has then been compared with the deviance of a model containing: only a random intercept for institution (two-level: institution); only a random intercept for department (two-level: department) and random intercepts for institution and department (three-level: institution and department). A lower deviance indicates a better model fit.

60. The variance parameter estimates for each of the random effects have also been considered. A statistically significant estimate means that there is evidence that the variation between units within the random effects (so between departments or institutions) differs.

61. Table 1 shows that in all of the models with random effects included, the variance parameter estimates are statistically significant at the α=0.001 level. This means that there is very strong evidence of difference in variation in the proportion of students achieving a first or upper second class degree between institutions, between departments, and between departments within institutions. In other words, some institutions and some departments perform differently to others in terms of the proportion of students achieving a first or upper second class degree. The deviance is lowest for the three level model, with random intercepts for institutions and departments within institutions, indicating that this model is the best fit for the data.

**Table 1: Results for modelling with no fixed effects**

| | | Null | Two-level: institution | Two-level: department | Three-level: institution and department |
|---|---|---|---|---|---|
| *Fixed effects* | | | | | |
| intercept | Estimate | 0.676*** | 0.496*** | 0.491*** | 0.456*** |
| | Standard error | 0.006 | 0.056 | 0.022 | 0.053 |
| *Random effects* | | | | | |
| Between department variance | Estimate | | | 0.495*** | 0.223*** |
| | Standard error | | | 0.025 | 0.014 |
| Between institution variance | Estimate | | 0.351*** | | 0.282*** |
| | Standard error | | 0.052 | | 0.046 |
| Deviance (-2 log-likelihood) | | 176798.4 | 169915.7 | 167649.9 | 167175.0 |

Note: ***=significant at the α=0.001 level.

62. The effect for department within institutions is much larger than the effect of institution alone, suggesting that the differences in achieving a first or upper second class degree are more

likely to be driven by differences within departments than by some behaviour displayed across the entire institution.

63. Since both random intercepts in the three-level model are found to be statistically significant, only the results for the three-level model will be reported on. Table 2 shows the estimates for all four three-level models that have been carried out:

    - Model 1: fixed effect for unadjusted tariff points only
    - Model 2: fixed effect for rounded adjusted tariff points only
    - Model 3: fixed effects for unadjusted tariff points and student characteristics
    - Model 4: fixed effects for rounded adjusted tariff points and student characteristics.

64. The deviance statistics show that Model 4 has the best fit, which is the model that uses the rounded adjusted tariff points and includes fixed effects for student characteristics. The fit is shown to be better for this model than for the model which includes the unadjusted tariff points, despite the estimates appearing to be equal for the different types of tariff points in the two models. The estimate for the rounded adjusted tariff points in Model 4 is marginally bigger than the estimate for the unadjusted tariff points in Model 3, although this difference is very small.

65. The variance parameters for the random intercepts also echo what was seen in Table 1, where the effect on the between-department variance is seen to be larger than the between-institution effect. Although, this difference becomes much clearer when prior attainment is also in the model.

**Table 2: Results for models 1-4**

| | | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|---|
| *Fixed effects* | | | | | |
| Intercept | | -2.959*** | -2.885*** | -2.883*** | -2.840*** |
| Unadjusted tariff points | | 0.012*** | | 0.012*** | |
| Rounded adjusted tariff points | | | 0.012*** | | 0.012*** |
| Sex | Female | | | 0.344*** | 0.364*** |
| Sex | Male | | | Reference | |
| Disability status | In receipt of Disabled Students' Allowance | | | -0.049 | -0.038 |
| Disability status | Disabled | | | -0.303*** | -0.299*** |
| Disability status | No reported disability | | | Reference | |
| Ethnicity | Asian | | | -0.412*** | -0.409*** |
| Ethnicity | Black | | | -0.641*** | -0.627*** |

| | | | | | |
|---|---|---|---|---|---|
| Ethnicity | Mixed | | | -0.339*** | -0.334*** |
| Ethnicity | Other | | | -0.461*** | -0.450*** |
| Ethnicity | Unknown | | | -0.411*** | -0.416*** |
| Ethnicity | White | | | Reference | |
| POLAR4 | Quintile 1 | | | -0.101*** | -0.096*** |
| POLAR4 | Quintile 2 | | | -0.050* | -0.048* |
| POLAR4 | Quintile 3 | | | -0.025 | -0.022 |
| POLAR4 | Quintile 4 | | | 0.013 | 0.016 |
| POLAR4 | Unknown | | | -0.362* | -0.362* |
| POLAR4 | Quintile 5 | | | Reference | |
| *Random effects* | | | | | |
| Between department variance | | 0.286*** | 0.299*** | 0.271*** | 0.282*** |
| Between institution variance | | 0.038*** | 0.044*** | 0.028*** | 0.032*** |
| Deviance (-2 log-likelihood) | | 159678.6 | 159498.8 | 158244.0 | 158006.4 |

Note: *= significant at α=0.05, **=significant at α=0.01, ***=significant at α=0.001

66. Models 3 and 4 include student characteristics – the contextual variables discussed in Section 2.5. From the models, the predicted probability for each individual has been calculated and then the mean probability of achieving a first or upper second class degree at a given institution has been calculated. Likewise, the proportion of students within an institution achieving a first or upper second class degree has also been calculated. The difference between these two probabilities is the value-added.

67. For Model 4, the value-added at the institution level varies from -0.163 to 0.180. That is, at the institution with the lowest value-added, the proportion of students achieving a first or upper second class degree is 16.3 percentage points lower than the model predicts. At the other end, at the institution with the largest value-added the proportion of students achieving a first or upper second class degree is 18 percentage points higher than the model predicted. The full list of institutional effects from Model 2 and Model 4 can be found in Annex B. However, institutions have been anonymised, since this is an experimental methodology.

68. To consider the impact of including contextual variables, the institutional value-added from Model 2 and Model 4 have been correlated. This has shown that there is a very strong positive relationship between the two sets of value-added ($\rho=0.974$, $\alpha<.001$). This indicates that, in this case, the decision to include or omit the contextual variables has little impact on the value-added calculated.

## 4.2. Pairwise comparison technique

69. This methodology is based on that used by Hoxby[26] for modelling value-added at selective schools in the USA in the paper "Computing the value-added of American post-secondary institutions" (where 'schools' are equivalent to English universities). Hoxby identifies two sources of bias in value-added estimates:

    a. Institutions select students from the pool of applicants (vertical selection).
    b. Students select the institution they will attend from the offers they have received (horizontal selection).

70. The process of selection by students and by institutions causes non-random assignment of students to institutions, which prevents simple comparisons of outcomes. Selection may occur on both observable and unobservable characteristics of both the students and the institutions. The multilevel model already described was used to account for selection on observable characteristics. Hoxby's pairwise methodology aims to account for selection on unobservable characteristics.

71. To overcome the two sources of selection bias, Hoxby attempts to construct vertical and horizontal 'experiments' in which students with 'identical' entry profiles attend different institutions. For this analysis, students are paired based on their SAT scores. These pairs of students can then be described as being quasi-randomly assigned to institutions.

72. The vertical experiments seek to identify a set of students among whom the institution chooses to make offers for reasons that are unrelated to their likely value-added. That is, for reasons other than their likely achievements at university given their prior attainment and predicted attainment. Admissions staff appear to make offers in a random fashion. The decisions will not actually be random, but could be based on factors which may be less likely to affect students' learning, such as extra-curricular activities. Because the students within this group are quasi-randomly assigned to institutions, their outcomes at different institutions can be directly compared to estimate value-added.

73. The horizontal experiments seek to identify a set of institutions among which students choose for reasons that are unrelated to their likely value-added. For example, students may be influenced by the weather on the day they visit the campus or where their friends are hoping to study. Because the institutions within this group are quasi-randomly chosen by students, their outcomes at different institutions can be directly compared to estimate value-added.

74. Each of the paired comparisons from the horizontal experiments will have vertical selection bias and vice versa. The unbiased estimates of value-added for each institution can be recovered by finding the institution-level fixed effects that best explain the combined results of the paired comparisons. The drawback of this approach is that it only enables estimation of value-added for students who are equally likely to be accepted or rejected by selective institutions.

[26] See Hoxby, C, 2015, 'Computing the Value-Added of American Postsecondary Institutions'. Available from: https://www.irs.gov/pub/irs-soi/15rpcompvalueaddpostsecondary.pdf [PDF]

75. The challenge of recreating this methodology is to construct groups from UK data that meet these criteria. The approach taken by Hoxby is, in part, driven by the data available in the USA. This is quite different to the UK data, mostly because the process of admissions is very different in the USA and the UK. The core problem is that the methods depend on the available data. Hoxby has the advantage of continuous prior attainment information which is not available in the UK. However, the UCAS data provides information on offers and acceptances which is not available in the USA. These differences in the data necessitate adapting Hoxby's methods in order to apply them to UK data.

76. The analysis is conducted at the group-level of student types (students entering higher education in the same year with the same prior attainment and applying to the same higher education institutions) who then go on to enrol at a particular higher education institution of interest. Hoxby uses the test score data submitted to colleges for submissions and can identify where students enrolled, but not the institutions where they were accepted but chose not to attend.

77. A further difference is that the measures of start and end point used by Hoxby, test scores and earnings, could be viewed as continuous variables whereas degree classifications are categorical. This will have an impact on the methodology, but is unlikely to be of any great consequence since the modelling technique uses neither the start or end point measures but relies on the outcomes from the pairwise comparisons, which are reliant on the start and end point measurements.

### 4.2.1. Vertical selection (institutions select from the pool of applicants)

78. Vertical selection is addressed using the fact that, typically, selective colleges in the US have institution-wide boundaries for test score acceptance or rejection, which enables Hoxby to identify institutional 'bubble ranges' where applicants within that range of test scores are equally likely to be accepted or rejected by the institution. However, Hoxby uses enrolment rather than admission outcomes for this assignment due to data availability.

79. Once the bubble ranges are identified, the applicants within this range are assumed to be randomly admitted or rejected. Therefore, the difference in outcomes can be calculated directly for each observed student type, at pairs of higher education institutions.

#### Issues with applying vertical selection method to the UK system

80. A-level entry information is not as granular as the US test scores. This means that identification of a bubble range of prior attainment scores is much harder: Further, the admissions behaviour at UK institutions is highly variable across different faculties and departments within the institution. Therefore, it is possible that the bubble range (the range of grades where applicants are equally likely to be accepted or rejected) could change across the institution, thus making the identification of the range difficult.

81. We have looked for equivalent ranges for English institutions by looking at tariff points, but have not identified similar patterns. This element of the methodology relies on the assumption that where students are in this situation, institutions quasi-randomly assign students to the 'accept' or 'reject' groups. Without identifying such a group in our data, we cannot make the same assumptions.

82. In order to address this, we have identified groups of students who fit the description of the bubble range. For each institution, the proportion of students given an offer for that institution who had a particular number of tariff points has been calculated. For any tariff points where the proportion falls between 40 and 60 per cent, all students with these results have been included in the analysis. This range has been chosen since restricting to only those where precisely 50 per cent of students with a particular number of tariff points gave very few results.

### 4.2.2. Horizontal selection (students select from institutional offers)

83. Hoxby assumes that students select colleges randomly within indifference sets (those sets of institutions where students are indifferent to which one they attend). That is, if the institutions the student applies to are all of the same selectivity and the student is likely to be admitted to all institutions, then the student randomly selects where to enrol.

84. Pairs of equally selective institutions are identified by comparing 25th and 75th percentiles of the maths and verbal test scores for enrolled students. Students very likely to be admitted are identified by selecting those with scores between the 65th and 80th percentiles of that institution pair. The analysis compares the difference in outcomes of students who went to Institution A with the outcomes of similar student types who went to Institution B with equal selectivity and where applications were made to both by those students.

#### Issues with applying horizontal selection method to the UK system

85. While the horizontal method can be applied to the UK system, Hoxby makes some assumptions which we do not need to apply with the UK data. The US data does not allow Hoxby to see whether applicants received offers from institutions, only their SAT scores, where they applied and where they attended. This means additional measures are put in place to support the quasi-randomisation assumption.

86. The additional checks ensure that the institutions compared are as selective as each other (to ensure institutions are not included in a set as ‘safety schools’) and that the students in the comparisons are very likely to attend (this is a proxy for not knowing if they actually received an offer to study). The UK data we hold has the advantage of containing prior attainment, application and offer information (including whether applicants received offers and whether the applicant chose the institution as their firm or insurance choice), attendance information and the outcome of that instance of study

87. When calculating the indifference sets at the institution level there is not much variation in the entry tariff point distributions for different institutions. Hoxby creates indifference sets of institutions, which are sets of institutions that students quasi-randomly select between. Colleges are considered to be ‘similar’ based on the average scores of their entrants (based on their selectivity). Hoxby does this by looking at institutions that have similar 25$^{th}$ and 75$^{th}$ percentile scores in both the maths and verbal tests. We attempted a similar approach based on entry tariff points, but found that there is not sufficient variation in the distribution of points across institutions for this to work. Therefore, we have created indifference sets by calculating the median tariff points for each institution then grouping institutions based on this.

### 4.2.3. Methodology

88. The number of students included in the pairwise methodology starts the same as for the multilevel methodology, as described in paragraph 51.These students are then reduced to only those who have admissions information in the UCAS data, leaving 126,520 students.

#### Vertical

89. For the vertical element, the number of students with a particular number of tariff points who applied to each institution was obtained. Then, the proportion of those who applied with that many tariff points and were offered a place was calculated. Where that proportion is between 40 per cent and 60 per cent, students applying to that institution with that many tariff points are included in the pairwise comparisons (i.e. these students are considered to be in the aforementioned bubble range).

90. For example, 30 students holding 200 tariff points applied to institution X, and 15 of them were offered a place. This would mean that the probability of being offered a place at institution X if you have 200 tariff points is 50 per cent. Therefore, students applying to institution X with 200 tariff points are included in the pairwise comparison.

91. Students have been found to be in the bubble range (that is, to have between and 40 and 60 per cent chance of being offered a place – as described in paragraph 78) at 39 institutions. This means that there are only 39 institutions in the analysis where between 40 and 60 per cent of students applying to that institution with a particular number of tariff points are offered a place at the university.

92. The outcomes of these students are then considered. The proportion of students attending the institution in question with a particular grade profile who achieve a first or upper second class degree is calculated, as is the proportion of students who applied to this institution with that same grade profile but who attended a different institution. The difference in these proportions is calculated for each combination of grade profile, institution applied to and institution attended.

93. In all, 4,333 comparisons are made – where the outcomes of students applying to and attending institution X with Y tariff points are compared with the outcomes of students applying to institution X with Y tariff points, but attending institution Z. These comparisons form the vertical element of the pairwise modelling.

#### Horizontal

94. For the horizontal element, we must identify students with the same number of tariff points who applied to two institutions within the same ‘indifference set’. Institutions within an indifference set should have a similar level of selectivity – or in other words, should, on average, select students with similar tariff points. To find indifference sets in the UK data, the median tariff points of those who started attending each institution in 2009 has been calculated. Institutions with the same median points have then been grouped into an indifference set. This created 10 indifference sets. However, one set contained only one institution, so this institution has been grouped with institutions whose median points is slightly higher. The indifference set with the highest median contained only the University of Oxford and the University of Cambridge. Since it is not possible to apply to both of these institutions, they have both been moved into the indifference set below.

95. For groups of students applying to institutions within an indifference set with the same grade profile, comparisons are then made of the proportion achieving a first or upper second class degree, and the difference is calculated between the proportion achieving well who attended institution X with Y tariff points and the proportion achieving well who applied to institution X with Y tariff points but attended institution Z. This appears identical to the vertical methodology, but here this comparison is only made where students apply to and attend institutions in the same indifference set, and there is no restriction based on the likelihood of being offered a place at the institution.

96. There are 108 institutions included in the horizontal element of this methodology and 14,859 comparisons made.

*Combining vertical and horizontal*

97. Having found the pairwise comparisons for the vertical and horizontal experiments, these have then been combined to give the value-added measure at the institutional level. The combined data contains 109 institutions.

98. To combine the data, dummy variables are created for each institution. These dummy variables indicate whether the students in a comparator group applied to that institution but did not attend, applied to that institution and did attend, or did not apply to the institution. However, dummy coding in this way removes the detail of the proportion of students achieving a first or upper second class degree and the number of students in each comparator group. This information is therefore included in the model as an observation weighting.

99. We used standard statistical equations to estimate the variance of the calculated difference in proportions for attended and comparator institutions. The inverse of these are then used to weight the model estimates, which weights model inputs dependent on the size of the population used to calculate the proportions.

100. The weight is calculated as 1 + the standard deviation of the difference between sample proportions ($\sigma_d$), which is approximately equal to:

$$\sigma_d = \left(\left[\frac{P_1(1-P_1)}{n_1}\right] + \left[\frac{P_2(1-P_2)}{n_2}\right]\right)^{\frac{1}{2}}$$

Where:

$P_1$ is the proportion of students applying to and attending institution X with Y tariff points who achieved a first or upper second class degree.

$n_1$ is the number of students applying to and attending institution X with Y tariff points

$P_2$ is the proportion of students applying to institution X with Y tariff points but attending institution Z who achieved a first or upper second class degree

$n_2$ is the number of students applying to institution X with Y tariff points but attending institution Z

101. Therefore, the model shows the difference in proportion of students with the same grade profile achieving a first or upper second class degree is equal to the vector of dummy variables indicating which institutions are being compared (X and Z), and the model is weighted by $1/\sigma_d$.

102. The model estimates, therefore, are the predicted percentage point difference in students who attended institution X achieving a first or upper second class degree compared to if they had attended a different institution. These estimates have been found to be statistically significant at the α=0.05 level for only 37 of the 109 institutions in the model. Of those, the estimates range from -0.259 to 0.167.

103. This does not seem to be a measure with an intuitive interpretation. In the case of the lowest estimate the model predicts that there will be a 25.9 percentage point difference in the proportion of students achieving a first or upper second class degree and students with the same entry profile who attend a different institution implying students attending that institution are less likely to achieve a first or upper second class degree than those who go elsewhere.

104. A full list of anonymised institutions and their estimated value-added can be found in Annex B.

# 5. Comparing the multilevel and pairwise techniques

105. Having calculated value-added using the two different approaches, it is of interest to assess how similar the estimates are. To do this, the value-added estimates from the two-level multilevel model with only rounded adjusted tariff points as a fixed effect have been correlated with the value-added estimates from the pairwise comparison method. The two-level multilevel model has been used (with a random intercept for institution only) since subject of study has not been included in the pairwise comparison method. The decision has been made to not include the contextual variables as they are also not included in the pairwise comparison method. The correlation shows that there is a moderate positive relationship between the value-added estimates from the two models ($\rho$=0.632, p<.0001). However, this is unsurprising since the methodologies are so different and the estimates are measuring value-added in different ways.

106. As well as considering the relationship between the two estimates, it is also of interest how closely the two sets of estimates rank institutions based on their value-added. To do this, the estimates have been ranked, lowest to highest, and then these ranks have been correlated. The ranks were also found to have a moderate positive relationship ($\rho$=0.623, p<.0001). Scatter plots for the method correlations can be found in Annex C.

107. This does not tell us that one model is better than the other, but it does tell us that the value-added that they are estimating is different. Since there is no standard measure of value-added to compare these estimates to, there is no straightforward way of assessing which is 'better'.

## 5.1. Strengths and weaknesses of the multilevel approach

108. The multilevel approach is a much studied and well-understood technique which is widely used in making institutional comparisons. It allows the hierarchical structure of the data to be accounted for and does not restrict the amount of data we are able to utilise (outside the restrictions of who to include that we have imposed). This means it is possible to include much more information than in the pairwise comparison method, such as subject details and student characteristics, because we are not restricted by small sample sizes.

109. However, the multilevel model can only account for observable factors – elements of the student and course characteristics which are already recorded in the data. This means that there will always be unexplained variance in the probability of achieving a first or upper second class degree.

110. At present, tariff points (both unadjusted and rounded adjusted) are entered into the model in a linear way. Better approaches might be to enter them as a polynomial, or as dummy variables, treating the variable as categorical rather than continuous, since the scores are measured in steps of 20.

111. Using the multilevel models in this analysis as a framework, it would be possible to expand the analysis to include a wider student population. This would not be without its challenges, such as finding a fair way of comparing all types of entry qualifications and deciding how to

assess outcomes from postgraduate qualifications if the population was extended to include all levels, but it is not the modelling approach that would make this difficult.

## 5.2. Strengths and weaknesses of the pairwise comparison approach

112. The pairwise comparison approach is a much more experimental approach than the multilevel methodology. It has not previously been applied to English institutions and the differences between the higher education system in the USA and the UK raise questions as to whether the way in which we have adapted the methodology may violate some of its underlying assumptions. Additionally, this approach is far less transparent than that multilevel method, since it is less familiar.

113. However, this approach does have the ability to account for unobservable factors by pairing students with similar grade profiles who make similar choices regarding which institutions to apply to. This is something that the multilevel model does not do. But these comparisons come at a cost to the amount of data we are able to use. Restricting to groups of students with the same grade profile who are either equally likely to be accepted or rejected at an institution they have applied to, or who have made applications to similar institutions, means that we are comparing small numbers of students with one another in many cases. This, in turn, prevents us from including any observable characteristics in the methodology to better assess whether students are 'similar', because this would result in groups of students that are too small to draw any fair comparisons.

114. Additionally, since the pairwise comparison method relies on admissions data, it will never be possible to expand this methodology to the whole student population. This is because not all institutions require undergraduate students to apply via UCAS, and the majority of postgraduate courses do not use UCAS at all.

## 5.3. Challenges for both methodologies

115. In reality, neither methodology is perfect. In the approach we have taken, both methodologies are restrictive in the students who can be included, excluding large parts of the student population. Even if we were to consider undergraduate students only, these current approaches are not able to consider students with unclassified degrees, which will remove the majority of students studying medicine or veterinary sciences, along with many students taking integrated masters courses. Likewise, they are only able to consider students with three or more A-Levels prior to starting higher education, which again greatly reduces the number of students included in this methodology.

116. Even having decided that it is acceptable for the framework to be developed using this restricted population, we still face a problem with both methodologies: how can we validate them? Since there is no standard approach to measuring value-added, how do we assess how well these methods are measuring value-added? It is certainly the case that both methodologies could have more sensitivity analyses carried out in order to assess how vulnerable they are to change if we change some of the choices we have made. This could, potentially, increase the confidence we have in the techniques, but if the two models still gave very different results, we still would have no way of knowing which approach is 'better'.

117. The biggest challenge of all, however, is that HELGA not only sought to measure value-added, but also to assess whether this could be used as a proxy for learning gain. Through this project, we have not been able to find a way of testing if this is the case.

# 6. Evaluating HELGA

118. HELGA set out to explore whether administrative data could be used to create a proxy measure for learning gain. The project has experimented with different techniques, considered different outcomes that could be a proxy for learning gain and different source of data that could be used. Ultimately, the project has developed two methods for measuring value-added in higher education for a subset of the undergraduate population. The two measures have produced different results in their measure of value-added and there is no straightforward way of evaluating which is the most accurate.

119. It should be noted that the 'value-added' measured is the difference between the 'expected' degree outcomes for students at an institution based on prior attainment (and other student and course characteristics in the case of the multilevel model) and the actual degree outcomes. The measure does not explain what this difference might be caused by. As mentioned in Section 2.3, there is concern about the comparability of degree classifications across institutions. This raises a question of the suitability of this value-added measure for comparing institutions.

120. Neither methodology should be used further without additional sensitivity analyses and serious thought as to what is really being measured and whether it is fair to measure institutional performance in terms of value-added based on the restricted population used.

121. At the outset of HELGA, it was known that it would never be possible to create a single measure of learning gain, encompassing all of the different elements that are understood to make up learning gain. It has necessarily focused on cognitive gain only, although some thought was given to using NSS metadata to measure non-cognitive learning gain, but this was unsuccessful. However, this does not mean that this could not be useful for measuring value-added, but it should be made clear that it should not be adopted to produce a single measure of learning gain.

122. There are a number of other elements that could be explored in this area of research that have been outside the scope of this project. These include:

a. Considering the impact of using different outcome measures

b. Seeing if it is possible to create a multi-dimensional measure of outcome that would allow for a more accurate measure of value-added (or indeed, learning gain)

c. Exploring whether use of GCSEs instead of A-levels, or some combination of the two, would affect the value-added estimates

d. Exploring whether it would be possible to use propensity score matching as part of the pairwise comparison method to give a tighter definition of students who are 'the same'

e. Considering how the work can be expanded to include more of the student populations, including qualifications on entry outside of A-Levels and levels other than undergraduate.

121. HELGA has in no way exhausted the potential avenues of research into the possibility of using administrative data for creating a proxy measure of leaning gain. However, it has shown that doing so is complex for many different reasons. It seems highly unlikely that a ‘one-size-fits-all’ measure of learning gain could ever be created from the administrative data. This is partly due to the fact that this data was not collected with the intention of measuring learning gain, but largely due to the complexity of the concept of learning gain.

# List of abbreviations

| | |
|---|---|
| **DfE** | Department for Education |
| **DLHE** | Destinations of Leavers of Higher Education (survey) |
| **HEFCE** | Higher Education Funding Council for England |
| **HELGA** | Higher Education Learning Gain Analysis |
| **HESA** | Higher Education Statistics Agency |
| **LDLHE** | Longitudinal Destinations of Leavers of Higher Education (survey) |
| **NMMLG** | National Mixed Methodology Learning Gain (project) |
| **NSS** | National Student Survey |
| **OfS** | Office for Students |
| **POLAR** | Participation of local areas |

# Annex A: Adjusting UCAS tariff points

1. The table below shows the correlation coefficients for the subject estimates of each individual year included in the modelling for adjusting tariff points (see Section 3). It shows that there is a strong positive correlation between all years, all of which are statistically significant at the α=.001 level. This means that the estimates do not significantly differ between years so the four years can be put together for the analysis without any concern that any one year is having undue influence on the results.

**Table A1: Correlation coefficient of subject estimates for each year included in the adjusting tariff points analysis.**

| | | Estimate 2008 | Estimate 2009 | Estimate 2010 | Estimate 2011 |
|---|---|---|---|---|---|
| **Estimate 2008** | Correlation coefficient | 1 | 0.932 | 0.930 | 0.909 |
| | p-value | | <.0001 | <.0001 | <.0001 |
| | Number of observations | 124 | 124 | 124 | 124 |
| **Estimate 2009** | Correlation coefficient | 0.932 | 1 | 0.955 | 0.938 |
| | p-value | <.0001 | | <.0001 | <.0001 |
| | Number of observations | 124 | 126 | 126 | 126 |
| **Estimate 2010** | Correlation coefficient | 0.930 | 0.955 | 1 | 0.994 |
| | p-value | <.0001 | <.0001 | | <.0001 |
| | Number of observations | 124 | 126 | 129 | 129 |
| **Estimate 2011** | Correlation coefficient | 0.909 | 0.938 | 0.994 | 1 |
| | p-value | <.0001 | <.0001 | <.0001 | |
| | Number of observations | 124 | 126 | 129 | 129 |

2. The model was run for only those A-level subjects with more than 500 students, those with 1,000 students or more, and for all subjects. In preliminary analyses, two of the largest estimates are for general studies and critical thinking. This may be because one or other of these subjects is mandatory at some Key Stage 5 providers. Because of this, the model was also run with the original reduction to only those subjects with 1,000 or more students but with general studies and critical thinking also removed. Finally, the model has been run with no restrictions on the number of students in a subject.

3. The estimated coefficients for subjects for the four models have been correlated, the results for which are shown in Table A2. These were all found to all be highly correlated with one another ($\rho>0.99$, $p<.001$). To further assess the differences in the outcomes, the estimated coefficients for subjects that appear in all four models were ranked and the ranks have then been correlated (Table A3). This has also shown that the results in all four cases are very similar ($\rho>0.99$, $p<.001$). While this is the case, the estimates for some of the smaller subjects (less than 500) in the model including all subjects are very large, likely due to the small number of participants. Therefore, the model that only includes those subjects with 500 or more students will be used, to maximise the coverage without risking the small student numbers affecting the output.

**Table A2: Correlation coefficients for subject estimates for the four different models**

| | | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|---|
| **Model 1** | Correlation coefficient | 1 | 0.999 | 0.999 | 0.999 |
| | p-value | | <.0001 | <.0001 | <.0001 |
| | Number of observations | 120 | 120 | 118 | 120 |
| **Model 2** | Correlation coefficient | 0.999 | 1 | 0.999 | 0.999 |
| | p-value | <.0001 | | <.0001 | <.0001 |
| | Number of observations | 120 | 133 | 118 | 133 |
| **Model 3** | Correlation coefficient | 0.999 | 0.999 | 1 | 0.999 |
| | p-value | <.0001 | <.0001 | | <.0001 |
| | Number of observations | 118 | 118 | 118 | 118 |
| **Model 4** | Correlation coefficient | 0.999 | 0.999 | 0.999 | 1 |
| | p-value | <.0001 | <.0001 | <.0001 | |
| | Number of observations | 120 | 133 | 118 | 448 |

Note: Model 1 = Model for only subjects with 1,000 or more students.
Model 2 = Model for only subjects with 500 or more students.
Model 3 = Model for only subjects with 1,000 or more students with general studies and critical thinking removed.
Model 4 = Model with no restriction on the number of students in each subject.

**Table A3: Correlation coefficient for ranked subject estimates for the four models**

| | | Ranks for Model 1 estimates | Ranks for Model 2 estimates | Ranks for Model 3 estimates | Ranks for Model 4 estimates |
|---|---|---|---|---|---|
| **Ranks for Model 1 estimates** | Correlation coefficient | 1 | 0.999 | 0.997 | 0.999 |
| | p-value | | <.0001 | <.0001 | <.0001 |
| **Ranks for Model 2 estimates** | Correlation coefficient | 0.999 | 1 | 0.997 | 0.999 |
| | p-value | <.0001 | | <.0001 | <.0001 |
| **Ranks for Model 3 estimates** | Correlation coefficient | 0.997 | 0.997 | 1 | 0.997 |
| | p-value | <.0001 | <.0001 | | <.0001 |
| **Ranks for Model 4 estimates** | Correlation coefficient | 0.999 | 0.999 | 0.997 | 1 |
| | p-value | <.0001 | <.0001 | <.0001 | |

Note: Model 1 = Model for only subjects with 1,000 or more students.
Model 2 = Model for only subjects with 500 or more students.
Model 3 = Model for only subjects with 1,000 or more students with general studies and critical thinking removed.
Model 4 = Model with no restriction on the number of students in each subject.

# Annex B: Estimates of value-added for the two models

1. Table B1 contains the value-added estimates from the two techniques: multilevel modelling and pairwise comparison. Estimates from both the two-level and three-level multilevel models are shown, since the two-level model is used for comparison with the pairwise comparison method. 'Prior attainment only' refers to the model with a fixed effect for prior attainment, measured using the rounded adjusted tariff points. 'All contextual variables' is the model with rounded adjusted tariff points plus fixed effects for sex, ethnicity, disability and POLAR4. The institutions have been anonymised since this methodology is experimental.

2. Empty cells in the pairwise value-added estimate column are due to these institutions not being present in the pairwise comparison methodology.

**Table B1: Estimated institutional value-added for two-level and three-level multilevel models with fixed effects for prior attainment only, with all fixed effects, and from the pairwise comparison method.**

| Institution | 2-level multilevel | | 3-level multilevel | | Pairwise comparison |
|---|---|---|---|---|---|
| | Prior attainment only | All contextual variables | Prior attainment only | All contextual variables | |
| A | 0.092 | 0.067 | 0.099 | 0.073 | 0.084 |
| B | -0.009 | -0.009 | -0.015 | -0.016 | -0.028 |
| C | -0.092 | -0.044 | -0.085 | -0.045 | -0.080 |
| D | -0.048 | -0.054 | -0.050 | -0.057 | 0.051 |
| E | 0.049 | 0.020 | 0.055 | 0.026 | 0.055 |
| F | 0.013 | -0.005 | 0.022 | 0.002 | 0.047 |
| G | 0.093 | 0.077 | 0.100 | 0.082 | 0.106 |
| H | 0.079 | 0.103 | 0.089 | 0.108 | 0.125 |
| I | -0.023 | -0.040 | -0.016 | -0.035 | 0.005 |
| J | 0.040 | 0.054 | 0.044 | 0.055 | 0.073 |
| K | -0.034 | -0.034 | -0.050 | -0.053 | . |
| L | 0.096 | 0.104 | 0.098 | 0.103 | 0.114 |
| M | 0.077 | 0.072 | 0.069 | 0.063 | 0.040 |
| N | -0.041 | -0.036 | -0.048 | -0.044 | 0.055 |
| O | 0.003 | 0.008 | -0.003 | 0.000 | -0.047 |
| P | -0.098 | -0.083 | -0.081 | -0.070 | -0.183 |
| Q | 0.014 | 0.008 | 0.009 | 0.002 | -0.091 |
| R | -0.077 | -0.055 | -0.082 | -0.065 | -0.018 |
| S | -0.429 | -0.400 | -0.533 | -0.521 | . |
| T | 0.020 | 0.010 | 0.023 | 0.013 | -0.029 |
| U | -0.064 | -0.049 | -0.080 | -0.067 | 0.049 |

| | | | | | |
|---|---|---|---|---|---|
| V | -0.053 | -0.063 | -0.035 | -0.046 | -0.055 |
| W | 0.034 | 0.043 | 0.046 | 0.052 | 0.016 |
| X | 0.071 | 0.057 | 0.079 | 0.063 | 0.085 |
| Y | -0.017 | -0.033 | -0.008 | -0.026 | 0.016 |
| Z | 0.001 | 0.005 | 0.002 | 0.008 | . |
| AA | -0.014 | -0.007 | -0.019 | -0.014 | 0.000 |
| AB | 0.074 | 0.054 | 0.081 | 0.061 | 0.080 |
| AC | -0.025 | -0.011 | -0.013 | -0.005 | -0.004 |
| AD | 0.032 | 0.020 | 0.066 | 0.052 | 0.445 |
| AE | -0.016 | -0.014 | -0.024 | -0.023 | 0.060 |
| AF | 0.008 | -0.017 | 0.021 | -0.004 | 0.004 |
| AG | -0.003 | 0.004 | -0.011 | -0.006 | 0.148 |
| AH | 0.060 | 0.048 | 0.072 | 0.057 | 0.020 |
| AI | 0.017 | 0.002 | 0.025 | 0.009 | 0.023 |
| AJ | -0.011 | -0.013 | -0.015 | -0.018 | -0.003 |
| AK | 0.014 | -0.005 | 0.024 | 0.005 | 0.051 |
| AL | 0.060 | 0.036 | 0.073 | 0.048 | -0.063 |
| AM | 0.022 | 0.030 | 0.019 | 0.024 | . |
| AN | 0.099 | 0.087 | 0.127 | 0.118 | 0.280 |
| AO | -0.005 | -0.019 | 0.009 | -0.006 | -0.144 |
| AP | 0.028 | 0.028 | 0.026 | 0.023 | 0.054 |
| AQ | 0.062 | 0.033 | 0.069 | 0.039 | 0.062 |
| AR | 0.049 | 0.018 | 0.063 | 0.029 | -0.002 |
| AS | 0.063 | 0.033 | 0.069 | 0.037 | 0.113 |
| AT | -0.105 | -0.069 | -0.113 | -0.081 | 0.047 |
| AU | 0.026 | 0.034 | 0.035 | 0.038 | 0.077 |
| AV | -0.032 | 0.002 | -0.023 | 0.005 | 0.018 |
| AW | -0.035 | -0.051 | -0.035 | -0.052 | 0.075 |
| AX | 0.055 | 0.024 | 0.063 | 0.031 | 0.096 |
| AY | 0.029 | 0.010 | 0.030 | 0.010 | 0.078 |
| AZ | -0.001 | 0.004 | 0.012 | 0.013 | -0.011 |
| BA | 0.049 | 0.040 | 0.077 | 0.067 | 0.127 |
| BB | 0.029 | 0.010 | 0.043 | 0.023 | 0.054 |
| BC | -0.011 | -0.012 | -0.003 | -0.007 | -0.011 |
| BD | -0.028 | -0.017 | -0.036 | -0.028 | 0.005 |
| BE | -0.013 | -0.041 | -0.005 | -0.033 | 0.024 |
| BF | -0.018 | -0.011 | -0.019 | -0.014 | -0.014 |
| BG | -0.128 | -0.101 | -0.136 | -0.113 | -0.093 |
| BH | -0.019 | 0.008 | -0.008 | 0.014 | 0.032 |

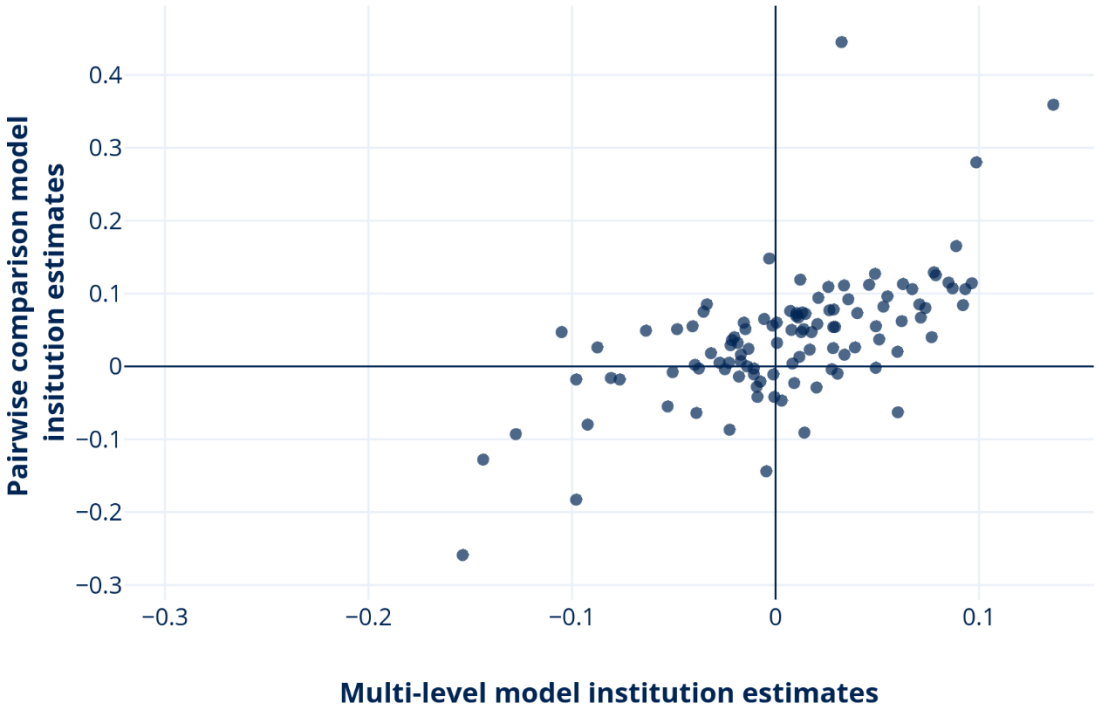
| | | | | | |
|---|---|---|---|---|---|
| BI | -0.001 | -0.013 | 0.006 | -0.008 | -0.042 |
| BJ | -0.040 | -0.057 | -0.030 | -0.048 | 0.002 |
| BK | -0.023 | -0.024 | -0.006 | -0.009 | -0.087 |
| BL | 0.021 | 0.070 | 0.020 | 0.062 | 0.094 |
| BM | 0.030 | 0.015 | 0.039 | 0.023 | -0.010 |
| BN | 0.011 | 0.025 | 0.018 | 0.028 | 0.067 |
| BO | -0.006 | 0.043 | -0.002 | 0.042 | . |
| BP | 0.039 | 0.035 | 0.035 | 0.029 | 0.026 |
| BQ | 0.046 | 0.054 | 0.043 | 0.049 | 0.112 |
| BR | 0.012 | -0.004 | 0.016 | -0.001 | 0.013 |
| BS | 0.067 | 0.041 | 0.071 | 0.043 | 0.106 |
| BT | -0.037 | -0.014 | -0.044 | -0.025 | . |
| BU | 0.007 | 0.020 | 0.002 | 0.012 | 0.076 |
| BV | -0.015 | -0.016 | -0.003 | -0.006 | 0.051 |
| BW | -0.002 | -0.012 | 0.008 | -0.005 | 0.056 |
| BX | 0.027 | 0.019 | 0.038 | 0.028 | -0.004 |
| BY | -0.007 | -0.028 | -0.004 | -0.027 | -0.021 |
| BZ | 0.026 | 0.031 | 0.046 | 0.048 | 0.109 |
| CA | -0.017 | -0.033 | -0.007 | -0.026 | 0.007 |
| CB | -0.020 | -0.035 | -0.022 | -0.041 | 0.040 |
| CC | 0.018 | -0.010 | 0.027 | -0.001 | 0.047 |
| CD | -0.144 | -0.143 | -0.156 | -0.163 | -0.128 |
| CE | -0.067 | -0.051 | -0.054 | -0.043 | . |
| CF | -0.022 | -0.020 | -0.030 | -0.029 | 0.036 |
| CG | 0.009 | 0.022 | 0.023 | 0.032 | -0.023 |
| CH | 0.008 | -0.012 | 0.019 | -0.001 | 0.050 |
| CI | 0.089 | 0.114 | 0.098 | 0.118 | 0.165 |
| CJ | 0.051 | 0.063 | 0.047 | 0.055 | 0.037 |
| CK | -0.039 | -0.042 | -0.061 | -0.073 | -0.064 |
| CL | 0.024 | 0.023 | 0.036 | 0.037 | . |
| CM | -0.034 | -0.048 | -0.034 | -0.055 | 0.085 |
| CN | 0.020 | 0.022 | 0.014 | 0.014 | 0.058 |
| CO | 0.001 | 0.002 | -0.003 | -0.004 | 0.032 |
| CP | 0.085 | 0.075 | 0.121 | 0.111 | 0.115 |
| CQ | 0.087 | 0.086 | 0.094 | 0.090 | 0.107 |
| CR | 0.028 | 0.017 | 0.034 | 0.022 | 0.025 |
| CS | 0.071 | 0.052 | 0.082 | 0.060 | 0.067 |
| CT | -0.038 | -0.009 | -0.041 | -0.017 | -0.003 |
| CU | -0.022 | -0.045 | -0.011 | -0.036 | 0.029 |

| | | | | | |
|---|---|---|---|---|---|
| CV | -0.051 | -0.044 | -0.056 | -0.051 | -0.008 |
| CW | -0.009 | -0.008 | -0.013 | -0.014 | -0.042 |
| CX | -0.006 | -0.004 | 0.003 | 0.001 | 0.065 |
| CY | 0.078 | 0.091 | 0.085 | 0.094 | 0.129 |
| CZ | 0.000 | 0.006 | -0.009 | -0.005 | 0.060 |
| DA | -0.154 | -0.173 | -0.132 | -0.150 | -0.259 |
| DB | -0.088 | -0.03 | -0.087 | -0.038 | 0.026 |
| DC | 0.136 | 0.122 | 0.190 | 0.180 | 0.359 |
| DD | 0.013 | 0.005 | 0.022 | 0.011 | 0.074 |
| DE | -0.098 | -0.068 | -0.105 | -0.080 | -0.018 |
| DF | 0.010 | 0.001 | 0.020 | 0.008 | 0.069 |
| DG | 0.010 | -0.018 | 0.018 | -0.011 | 0.073 |
| DH | 0.015 | -0.005 | 0.027 | 0.005 | 0.072 |
| DI | 0.053 | 0.044 | 0.060 | 0.050 | 0.082 |
| DJ | 0.012 | 0.006 | 0.014 | 0.006 | 0.119 |
| DK | 0.036 | 0.034 | 0.034 | 0.031 | 0.092 |
| DL | 0.034 | 0.031 | 0.041 | 0.036 | 0.111 |
| DM | -0.081 | -0.040 | -0.082 | -0.047 | -0.016 |

# Annex C: Comparing value-added estimates from the multilevel modelling and pairwise comparison techniques

Figure C1 below shows the correlation of the value-added estimates from the multilevel model containing random intercepts for institution only and fixed effects for student characteristics and the value-added estimates from the pairwise comparison method. As discussed in paragraph 105, this shows that the models are producing different estimates for institutions and while they are correlated, it is not a strong relationship.

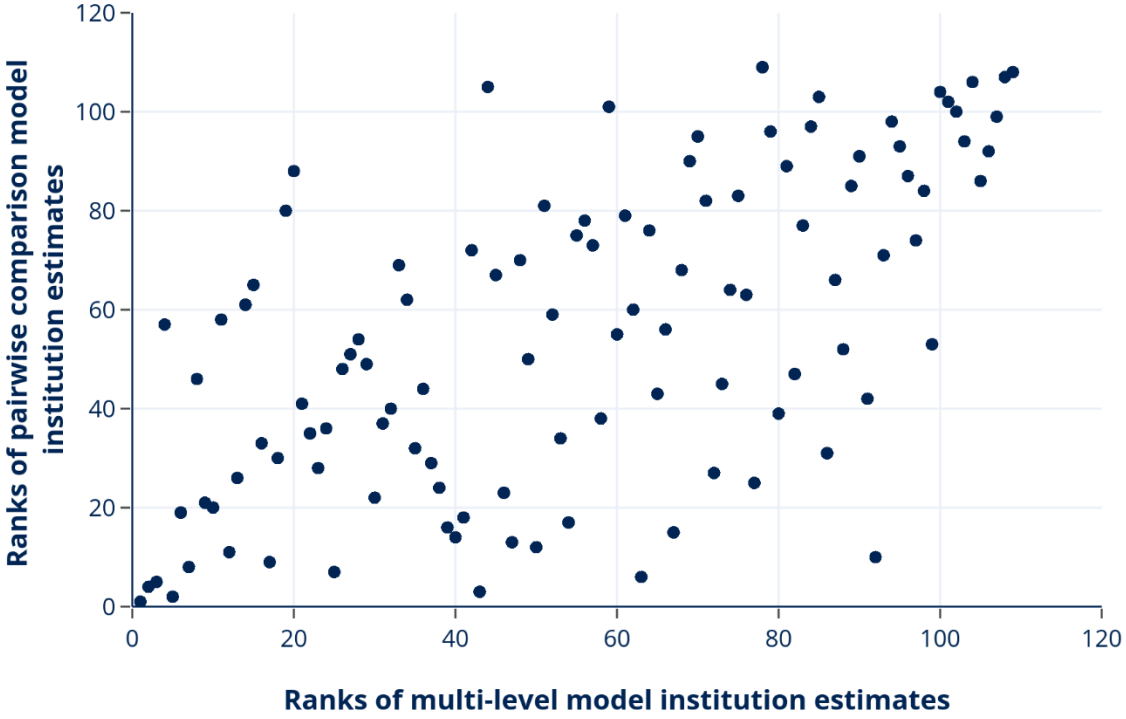
**Figure C1: Correlation of institution value-added estimates from the multilevel and pairwise comparison models**




Note: ρ=0.63, α < .001

It is also of interest to assess whether value-added estimates are ranking institutions similarly, or differently. To do this, the value-added estimates for the two models have been ranked, and those ranks have been correlated. The results can be seen in Figure C2 below.

**Figure C2: Correlation of ranked institution value-added estimates from the multilevel and pairwise comparison models**




Note: ρ=0.62, α < .001