



Department
for Education



Using graduate earnings data in the regulation of higher education providers

Research report

May 2025

Jack Britton, Kate Ogden and Ben Waltmann

Institute for Fiscal Studies



Government
Social Research

Contents

List of tables and figures	2
Executive summary	5
1. Introduction	8
Definitions	9
2. Earnings and the current regulatory framework	11
3. Criteria for a regulatory earnings metric	14
Conceptual criteria	14
Practical criteria	18
4. Data	23
5. Key analytical decisions and introducing our ‘baseline’ approach	28
6. Baseline estimates	42
Results for a single subject across providers	46
Results for a single provider across subjects	49
Integrating our results with the OfS dashboards	54
7. Assessing the baseline metric	60
Performance against our criteria for a regulatory earnings metric	60
Comparison with the existing OfS progression metric	64
8. Sensitivity analysis	69
9. Labour market participation	84
10. Recommendations	88
Appendix	99
References	118

List of tables and figures

Figure 1: Number of graduates by level, mode and academic year of graduation	24
Figure 2: Years of available data from different administrative data sources.....	25
Figure 3: Match rates between data sets for the 2015/16 graduation cohort.....	26
Figure 4: Median earnings by subject area and years after graduation	29
Figure 5: Share of graduates changing earnings rank by more than 20 percentiles	32
Figure 6: Buckets for which estimates can be provided by level and mode.....	41
Table 1: Coverage (all restricted to buckets for which we report at least one estimate) ..	43
Figure 7: Distribution of earnings metric by course for all level–mode combinations.....	44
Figure 8: Baseline benchmark and difference from benchmark estimates, all courses, full-time first degree undergraduates	45
Figure 9: Average benchmark and actual earnings for full-time first degree courses in history	47
Figure 10: Percentage difference from benchmark for full-time first degree students in history	48
Figure 11: Average benchmark and actual earnings for full-time first degree students at an example provider	50
Figure 12: Percentage difference from benchmark for full-time first degree students at an example provider	51
Table 2: Full results for an example provider.....	52
Table 3: Additional contextual information for an example provider.....	53
Figure 13: Overview of progression outcomes for an example provider, OfS B3 dashboard.....	55
Figure 14: Provider-level differences from benchmark at an example provider	57
Figure 15: Detailed progression outcomes for an example provider, OfS B3 dashboard, indicator values.....	57
Figure 16: Detailed progression outcomes for an example provider, OfS B3 dashboard, differences from benchmark	58
Figure 17: Distribution of standard errors by level–mode, course-level estimates	61
Figure 18: Actual average earnings and OfS progression indicator value, by course.....	65

Figure 19: Difference from benchmarks for baseline earnings metric and B3 progression indicator, by course.....	67
Figure 20: Standard errors on baseline earnings metric and on B3 progression metric, by course	68
Colour-coding for sensitivity checks.....	70
Table 4: Different options for pooling across cohorts.....	71
Table 5: Different options for transforming the earnings outcome	73
Table 6: Different options for the earnings outcome used.....	74
Table 7: Different options for prior attainment controls	76
Table 8: Different options for accounting for location when earning	78
Table 9: Different options for accounting for course-level characteristics	81
Table 10: Non-parametric approaches	83
Figure 21: Benchmark share earning less than the NLW against percentage difference from benchmark.....	86
Figure 22: Benchmark share earning less than £3,000 per year against percentage difference from benchmark	87
Colour-coding for policy options.....	92
Table 11: Comprehensive table of options considered	92
Figure 23: Median earnings by subject area and years after graduation: women.....	99
Figure 24: Median earnings by subject area and years after graduation: men	99
Figure 25: Provider-level differences from benchmark, full-time first degree	100
Figure 26: Benchmark earnings and OfS progression benchmark, by course	100
Figure 27: Percentage difference from benchmark and selectivity of course: (a) baseline earnings metric and (b) OfS progression indicator.....	101
Table 12: Control variables included in the baseline model.....	101
Table 13: Coverage with alternative sample restriction of 23 FPE students (all restricted to buckets for which we report at least one estimate)	104
Figure 28: Earnings metric for full-time first degree courses in history: different options for pooling across cohorts	105

Figure 29: Earnings metric for full-time first degree courses in history: different options for transforming the earnings outcome	105
Figure 30: Earnings metric for full-time first degree courses in history: different options for the choice of earnings outcome	106
Figure 31: Earnings metric for full-time first degree courses in history: different options for controlling for prior attainment	106
Figure 32: Earnings metric for full-time first degree courses in history: different options for accounting for location when earning	107
Figure 33: Earnings metric for full-time first degree courses in history: different options for controlling for course-level characteristics	107
Figure 34: Earnings metric for full-time first degree courses in history: non-parametric approaches	108
Figure 35: Distribution of highest earnings three–five years after graduation, full-time first degree graduates.....	108
Table 14: Full list of metrics used to examine methodological variations.....	109
Figure 36: Baseline benchmark and difference from benchmark estimates, all courses, full-time other undergraduates	111
Table 15: Decomposition of variance in actual earnings / indicator value.....	112
Figure 37: Distribution of standard errors on estimates, for baseline and alternative specifications	112
Table 16: Impact of different criteria and thresholds on courses flagged for regulatory action	113
Table 17: Impact of different methodological variations on courses for which difference from benchmark is below –10 percentage points.....	114
Table 18: Impact of different criteria and thresholds on courses flagged for regulatory action, by broad subject area.....	115
Figure 38: Proportion of students on courses that would be flagged as being below different numerical thresholds under our baseline approach	117
Figure 39: Difference from benchmark and proportion of students with any self-assessment (SA) earnings in their highest-earning year three–five years after graduation, by course	117

Executive summary

This report explores options for incorporating administrative records on the earnings outcomes of higher education graduates into the Office for Students' (OfS's) regulation of the higher education sector in England. It focuses on developing metrics that can effectively measure higher education providers' (HEPs') success in equipping graduates with valuable labour market skills.

The OfS currently uses three key metrics, known as B3 metrics, to assess providers' performance: continuation (whether students continue their studies), completion (whether students complete their qualifications) and progression (whether graduates move into managerial or professional employment or further study). These metrics are compared against absolute thresholds; provider-specific benchmarks are also reported but only serve as contextual information.

In developing a new earnings-based metric, we outline both conceptual and practical criteria. Conceptually, an ideal metric should reflect the impact a course has on graduates' earnings and capture only factors that providers can control. Practical considerations include fairness across providers, accuracy, creating positive incentives for providers, regulatory practicality, timeliness and transparency.

A key outcome from the work is that we do on balance recommend using an earnings metric calculated from administrative data in regulation. We believe that the best option would be to integrate this new metric into the current OfS regulatory framework in such a way that it complements – rather than replaces – the existing B3 progression metric. Should it be used, this new metric could enable the OfS to develop a more comprehensive view of providers' performance in preparing students for the labour market.

Unlike the current B3 progression metric, all the options we consider in detail would use the difference from an institution-specific benchmark as the headline measure, targeting the causal effect of a course on graduates' earnings relative to other courses in the same subject area. This is because, in our view, only this type of metric can fulfil our conceptual and practical criteria to a sufficient degree.

We organise our discussion of different options for an earnings metric of this type around a baseline approach and various potential variations of it. Variations we examine include different approaches to pooling cohorts, transforming earnings outcomes, controlling for prior attainment and demographics, accounting for location when earning, and controlling for course-level characteristics. While our baseline estimates prove robust to many of these variations, some alternative specifications yield noteworthy differences. These alternatives provide policymakers with a range of options to consider when implementing an earnings metric.

Our proposed baseline metric is based on the highest earnings of graduates three to five years after graduation, although using data anywhere from two to six years after graduation, and averaging rather than taking the maximum, would be perfectly reasonable in our view. We recommend pooling data across two or three cohorts of graduates to maximise statistical power. Crucially, the metric would control for prior attainment and various demographic characteristics as well as subject studied, in line with the OfS's current approach to benchmarking.

We recommend estimating the model excluding very low earners, as very low recorded earnings are unlikely to reflect graduates' true earnings potential. Our baseline approach is to exclude those whose maximum earnings three to five years after graduation are below £3,000 in all years. To be transparent about the share of graduates excluded by this threshold, we recommend reporting this share alongside our main estimates for each course. If a much higher minimum earnings threshold such as full-time earnings on the national living wage was chosen, we would recommend estimating separate models for the share of graduates earning below (or above) that threshold, which could be adopted as a third progression indicator.

Our analysis reveals a persistent relationship between course selectivity and earnings outcomes across most specifications. We examine what we believe to be credible alternative approaches to our baseline approach that would remove this correlation, which could be considered by the regulator. However, as the assumptions underlying these approaches are strong, our recommendation is instead to use one of these approaches to generate an additional 'selectivity-adjusted benchmark' to be reported alongside our main benchmark.

We recommend not to control for graduates' location when earning, as enabling geographic mobility is one channel through which a course may impact graduates' earnings. Controlling for this would therefore move the earnings metric away from measuring the impact of a course on the earnings of its graduates. Instead, we recommend controlling for graduates' home region and reporting contextual information about graduates' location decisions separately, such as the share of graduates staying in the same local area as the provider or the share moving to London. However, we acknowledge that making an adjustment for the location of a provider's graduates may be desirable from a policy perspective. If so, we would recommend adjusting earnings for regional price levels, though this would require robust regional price indices that are yet to be developed.

We advise caution in reporting earnings metrics for part-time and Other Undergraduate courses. Estimates for these level–mode combinations will be substantially less robust than those for first degrees, partly due to the much lower number of students taught, and partly due to the larger share of mature students for whom school records are unavailable. In our view, there is therefore a good case not to proceed with earnings metrics for these level–mode combinations at present.

Importantly, any earnings metric should be treated as one of many sources of evidence on provider performance. Equipping graduates with skills that are valuable in the labour market is only one aspect of course quality. Earnings also only capture one aspect of graduates' success in the labour market; other factors such as job security or fulfilment at work may be as or more important for graduates. No earnings metric could fully account for the complexity of the higher education landscape. We recommend that the OfS continue considering the wider context in individual cases.

Our recommendations aim to balance various practical and conceptual criteria, providing a framework for implementing an earnings metric that could be useful to the OfS in assessing and regulating higher education providers. For a comprehensive list of recommendations and options considered, please refer to Section 10 of the report.

Acknowledgements

We gratefully acknowledge and thank members of our Advisory Group for their constructive input to this project. The Advisory Group included Professor Dame Alison Wolf (King's College London) and Nick Hillman (Higher Education Policy Institute), as well as representatives from the 10 Downing Street Data Science Team; the Association of Colleges; the Association of Employment and Learning Providers; the Department for Culture, Media and Sport; the Department for Education; Guild HE; HM Treasury; Independent Higher Education; the Office for Students; and Universities UK.

1. Introduction

The purpose of this report, which was commissioned by the Department for Education (DfE) in September 2023, is to set out options for making use of graduate earnings data to support the Office for Students' (OfS's) regulation of the higher education sector. Our aim is a measure of graduates' earnings outcomes that would be as informative as possible about higher education providers' (HEPs') success, or otherwise, in equipping graduates with skills that are useful for them in the labour market.

There are three main reasons to explore the use of earnings outcomes in the regulation of the higher education sector. The first is that the impact of a higher education course on graduate earnings will be an important aspect of a course from the point of view of many prospective students. Second, earnings outcomes are an indirect measure of course quality: all else equal, higher-quality courses will usually lead to higher graduate earnings. Third, unlike many other potential measures of course quality, graduate earnings can be measured objectively, accurately and with near-universal coverage.

However, there are also clear drawbacks. Other aspects of labour market progression, such as job security or fulfilment at work, may be as or more important for prospective students. This is especially relevant in subject areas such as creative or performing arts, where success as an artist may entail *lower* earnings than what the same person would have earned in a different occupation. A focus on earnings as an indirect measure of course quality could lead providers to focus on skills that are valuable in the labour market at the detriment of other aspects of course quality.

There are practical challenges as well. It will never be possible to completely disentangle differences in course quality from differences in the characteristics of a course's intake. Meaningful earnings data are only available several years after each cohort's graduation, so changes in course quality can only be picked up with a long lag. The advantages of an earnings metric for regulation need to be weighed against these drawbacks, along with the potential additional cost of implementing it.

The scope of this report is limited to undergraduate students, but similar metrics could also be developed for other levels for study. The options we have explored are designed to fit into the existing system of student outcome measures used in university regulation, which is described in Section 2.

The process of developing such an earnings metric can usefully be divided into two steps. The first step is to set out criteria by which a potential regulatory metric can be judged. These criteria may be conceptual or practical. Conceptually, it needs to be decided what the target or goal for a metric is; in other words, what the metric would ideally be measuring in the absence of data constraints. Practical criteria may relate, for instance, to the transparency of a metric or the ease of implementing it. Conceptual and practical criteria for evaluating a regulatory metric are set out in Section 3.

The second step is to assess how different possible metrics perform against these criteria. Section 4 presents an overview of the available data, which inherently limit the range of feasible metrics. In Section 5, we introduce our baseline metric methodology, which serves as a framework for subsequent discussions. Section 6 applies this baseline metric to the 2014/15 and 2015/16 graduation cohorts, presenting the results. Section 7 assesses the performance of our baseline metric against the criteria we established and compares it with the existing OfS progression metric. Section 8 conducts an extensive sensitivity analysis, exploring how our estimates change under various methodological choices. Section 9 considers options for incorporating labour market participation into the metric. Finally, Section 10 summarises our key recommendations.

Definitions

Bucket	We refer to each unique level–mode–subject combination as a bucket. For example, full-time first degree in mathematics would be one bucket.
CAH2	The second level of the Common Aggregation Hierarchy for grouping higher education subjects. This is the subject classification we use, which is also commonly used by the OfS and the Higher Education Statistics Agency (HESA). Throughout, we follow the OfS and aggregate Celtic studies (CAH19-02) with the languages and area studies grouping (CAH19-04).
Course	We refer to each unique level–mode–subject–provider combination as a course. This means that a HEP can have either one or zero courses in a given bucket.
First degree	HESA classification for standard three-year undergraduate degrees.
FPE	Full-person equivalent. Students graduating with joint-honours degrees will often have studied more than one CAH2 subject. An FPE number of students graduating in a particular subject or course is the number of students on that subject or course, where each student is counted by the proportion of time they spent on it.
HEP	Higher education provider. We use this more general term than ‘university’ because much higher education is provided at further education colleges.

Level	A graduate's 'level of study' according to HESA definitions, based on the qualification aim. The three levels we consider in this report are first degree undergraduate degrees, other undergraduate degrees (Other UG) and undergraduate degrees with postgraduate components (UG with PG).
Level-mode	It is sometimes convenient to use one term to capture all the different levels and modes of study. We use 'level-mode' for this, so that this would include 'full-time first degrees' and 'part-time first degrees' in the list, amongst others.
Mode	Either part-time or full-time.
Other UG	Other undergraduate degree. HESA classification for undergraduate degrees that do not count as first degrees.
UG with PG	Undergraduate degree with postgraduate components. HESA classification for undergraduate degrees that incorporate postgraduate elements. These are largely integrated master's courses.

2. Earnings and the current regulatory framework

The regulator for higher education in England is the Office for Students (OfS). Its four primary regulatory objectives are that all students:

1. Are supported to access, succeed in, and progress from, higher education.
2. Receive a high-quality academic experience, and their interests are protected while they study or in the event of provider, campus or course closure.
3. Are able to progress into employment or further study, and their qualifications hold their value over time.
4. Receive value for money.¹

To achieve these objectives, the OfS imposes '(ongoing) conditions of registration' on providers. These can be general conditions applying to all providers or specific conditions imposed on a particular provider. The general ongoing conditions of registration cover a range of requirements from access and participation plans to the payment of OfS registration fees.

The general condition of registration that relates to graduate outcomes is condition B3. Condition B3 requires that the provider must 'deliver positive outcomes for students on its higher education courses'. Delivering positive outcomes means that a provider performs at least as well as a numerical threshold set by the OfS on each of three indicators set by the OfS relating to *continuation* of studies, *completion* of studies, and *progression* into managerial or professional employment or further study.

A student is said to have a positive *continuation* outcome if they are still on their course (or have obtained a qualification) one year and two weeks after starting it. Similarly, a student is said to have a positive *completion* outcome if they have obtained a qualification (or are still on their course) four years and two weeks after starting it. A student is said to have a positive *progression* outcome if, 15 months after gaining their qualification, they are (a) in managerial or professional employment, as defined by the Office for National Statistics' Standard Occupational Classification; (b) in further study at any level of study; or (c) travelling, caring for someone else or retired. Continuation and completion indicators are taken from administrative data, whereas the progression indicator relies on student self-reports when they are contacted for the Graduate Outcomes survey.

For each indicator, minimum thresholds are set following a complex process that takes into account sector-level averages, regression analysis aiming to determine the maximum plausible effect of student intake composition on each outcome, and regulatory

¹ See page 13 of <https://www.officeforstudents.org.uk/media/fmzbr50j/securing-student-success-regulatory-framework-for-higher-education-in-england-2022.pdf>.

judgement. Thresholds vary across levels and modes of study but not across subjects, providers or courses.²

Each provider's performance on the three outcome indicators is assessed against the relevant threshold to determine whether a given provider is meeting condition B3. Performance is assessed overall and for various 'split indicators', which split the data for each indicator by subject, age on entry and various student background characteristics. Providers are meant to show performance at or above the threshold for all split indicators and overall to satisfy the condition. Where providers do not show performance at or above the threshold, they still count as satisfying the condition if contextual factors justify the observed outcomes or if there is an insufficient number of students for whom data are available.

One such contextual factor is a provider's performance relative to its OfS *benchmark*. The OfS estimates a benchmark for each indicator (overall and split) and provider. If a given outcome is below the numerical threshold for a provider but above its benchmark, this may count in a provider's favour. However, the OfS is clear that it will not treat a provider's benchmark values as determinate of whether it satisfies condition B3.

If a provider is found to be in breach of condition B3 or any other ongoing condition of registration, the OfS can intervene by (a) launching a formal dialogue with the provider and/or an investigation; (b) imposing enhanced monitoring requirements on the provider; or (c) imposing additional specific ongoing conditions of registration. In severe cases, it can also impose formal sanctions such as monetary penalties, a temporary suspension of the provider or, as a last resort, deregistration of the provider.

Administrative data on graduate earnings would most naturally fit into this framework as a replacement of or complement to the current B3 indicator on 'progression'. Compared with the current progression indicator, an indicator based on administrative data would likely be more reliable. Only 44% of graduates currently answer the Graduate Outcomes survey from which the progression indicator is derived, whereas 98% of graduates can be linked to a record in the administrative earnings data, with nearly 90% recorded as having non-zero earnings in the fifth full tax year after graduation. The lower response rate on the Graduate Outcomes survey can lead to bias in the current progression indicator, even conditional on observed characteristics, if those graduates who answer the survey are systematically different from those who do not. Using administrative earnings data would also make it possible to assess outcomes later in graduates' careers; as discussed further below, this would be desirable, given that it typically takes some years for graduates to 'find their feet in the labour market'.

However, as we argue below, it would not be desirable to simply slot a measure of average graduate earnings into the regulatory framework outlined above. This is

² See <https://www.officeforstudents.org.uk/media/7538/regulatory-advice-20-regulating-student-outcomes-revised.pdf>.

because, even more so than the current progression measure, average graduate earnings will be substantially affected by factors that are not directly influenced by the quality of their courses. The remainder of this report is concerned with developing a more appropriate earnings metric for regulation than graduates' average earnings relative to an absolute threshold.

It is worth noting that there is precedent for such a metric being calculated by the OfS: an earnings metric based on LEO data was used as a supplementary measure of graduate outcomes in the 'Year Three' (2018) and 'Year Four' (2019) Teaching Excellence Framework (TEF) assessments, as well as in the TEF subject pilot (Department for Education, 2017; Office for Students, 2018a and 2018b). Like other graduate outcome measures for the TEF, graduate earnings were evaluated relative to a provider-specific benchmark, which took into account both a provider's subject mix and the background characteristics of its intake. The TEF earnings metric thus followed the same broad structure that we propose for a regulatory earnings metric in this report.

Throughout the remainder of the report, we focus on developing an earnings metric for each *course* – by which we mean a unique combination of provider, level, mode and subject – as requested by the Department for Education. We agree with the Department that courses are the most important unit of analysis for an earnings metric, as we would expect providers to be able to affect earnings outcomes primarily at the course level. Note that this contrasts with the framing of the existing condition B3 metrics, which are provided primarily at the provider level at each mode and level of study, with a further split by subject provided as just one of many 'split indicators'. We discuss how provider-level estimates, as well as split indicators, could be constructed from our course-level estimates in Section 6.

3. Criteria for a regulatory earnings metric

This section sets out high-level criteria by which a regulatory metric or set of metrics for graduates' earnings outcomes may be judged. Two of these criteria are conceptual, i.e. they relate to what an ideal metric should be measuring in the absence of data and other practical constraints. Others are more practical, relating, for instance, to how transparent a metric would be to stakeholders or how it would fit into the existing regulatory framework.

Conceptual criteria

A key precondition for developing a regulatory metric for graduate earnings is a clear conception of what such a metric would ideally be measuring. The current regulatory framework for higher education provides some guidance: as set out in Section 2 above, the OfS's condition B3 requires that a provider deliver positive outcomes for students. In our view, this suggests two key conceptual criteria for evaluating the target of a graduate earnings metric: that it should reflect the *impact* a course has on the earnings of its graduates and that it should only capture the influence of factors that providers can *control*. While these two goals are largely compatible, a trade-off between them is unavoidable, as will become clear below.

The impact a course has on graduate earnings is a *causal effect*, i.e. a difference between graduates' actual outcomes and their own (unobservable) counterfactual outcomes had they not attended the course but done something else instead (see more on the 'something else' below). A causal target would capture a course's 'value-added' and thus avoid unfairly penalising providers whose students had lower earnings potential even before they enrolled on the course. This is in contrast to alternative metrics such as average earnings of students on a course, which penalise providers serving disadvantaged students almost by construction.

In the academic economics literature, the causal effect of higher education on a graduate's earnings is known as the return to higher education. A series of publications by the Institute for Fiscal Studies has provided recent estimates of the return to higher education in England, across subjects, institutions and courses (Belfield et al., 2018; Britton et al., 2020; Britton et al., 2022). Here we focus on estimates for courses, as this is the level of analysis that is most relevant for university regulation.

Broadly speaking, existing returns estimates for courses vary conceptually across three dimensions.

- **The time horizon over which graduate earnings are measured.** It seems clear that an ideal regulatory metric would consider earnings outcomes across the whole of a graduate's lifetime, as labour market success late in life will still

be of value for graduates. However, as we discuss below (under practical criterion v), there are obvious constraints on this in practice: graduates' earnings outcomes in their late careers will typically only be known half a century after graduation, at which point they are no longer useful for regulation.

- **Which group of students the estimates relate to.** The highest-quality academic evidence on returns to higher education courses is based on quasi-experimental methods, which exploit the near-random allocation of 'marginal' students across courses. However, returns for 'marginal' students may plausibly be different from returns for those students who comfortably got admitted to a given course. From a regulatory perspective, the object of interest is likely to be the average return across students who were actually enrolled on a given course.
- **The counterfactual against which graduates' earnings are compared.** This is likely the most important choice. The options include comparing against the counterfactual of not going to university ('absolute returns') and comparing against the counterfactual of studying a different course ('relative returns'). The advantages and disadvantages of these choices are discussed further below.

The main advantage of targeting the 'absolute return' of each course is that it is a very intuitive measure of the impact of a course on graduate earnings. However, an important drawback is that it conflicts with the criterion that, as far as possible, a regulatory metric should reflect providers' efforts rather than luck or other factors unrelated to providers' actions. Graduate earnings and thus absolute earnings returns will depend on the labour market value of skills in a particular subject area, which is determined by the demand for and supply of such skills in the wider economy rather than by any particular provider's performance.³ Because the absolute earnings return to a given course will also depend on graduates' counterfactual earnings had they not gone to university, it will also depend on factors influencing these counterfactual earnings such as minimum wage legislation or vocational training opportunities in students' local areas.

Targeting absolute returns in graduates' early careers with an earnings metric would be especially problematic when combined with a focus on early-career earnings, as much of the absolute return to higher education (HE) only arises in graduates' mid and late careers. This is because going to university means giving up on valuable labour market experience. In graduates' early careers, the returns to labour market experience in the non-HE counterfactual can mask the lifetime absolute return to higher education.

An alternative is to focus on 'relative' earnings returns, which compare graduates' earnings on a given course with the earnings that the same graduates would have

³ Prospective students learning about the earnings returns to different degrees may indeed be one factor affecting the supply of skills in a subject area and thus earnings returns in that subject area for future graduates.

received had they studied a different course. This would eliminate the influence of factors related to the non-HE counterfactual. Going further still, by comparing only with courses *within a given subject*, the effects of the labour market conditions for graduates of a given subject would also be eliminated. These changes to the relevant counterfactual move the target closer to something that providers can directly influence, while arguably still reflecting the impact of a course on graduate earnings relative to other courses in the same subject area.

In sum, even if a causal target were chosen, a policymaker would need to decide whether it was going to be (a) the causal effect of a course on its graduates' earnings relative to not attending higher education, (b) the causal effect of a course on its graduates' earnings relative to all other courses or (c) the causal effect of a course on its graduates' earnings relative to other courses in the same subject area. The most appropriate target may depend on what the regulator was seeking to do: (a) might be the most relevant metric when considering the overall shape of higher education provision, while (c) might be most appropriate when evaluating individual providers.

In the remainder of this report, we focus on a version of (c), the causal effect of a course on its graduates' earnings relative to other courses in the same subject area. To us, this is a reasonable compromise between the two criteria of accurately reflecting the impact on graduate earnings and of measuring the provider's actual efforts rather than extraneous factors. As set out further below, this approach also ranks highest on practical criteria.

Targeting (c) – or indeed (a) or (b) – would be a departure from the OfS's usual approach to regulation, which emphasises actual ('raw') indicator values relative to absolute thresholds that do not vary with a provider's intake (see Section 2). We do not recommend adopting the same approach for earnings, as average earnings will be substantially affected by factors that are not directly influenced by the quality of a given course. Such extraneous factors will include pre-university characteristics such as graduates' school attainment, but also labour market factors such as the supply and demand for graduates in a particular degree subject as well as graduate choices such as their choice of profession or their choice of whether to work part-time or full-time. As a result, actual average earnings of graduates in relation to an absolute threshold will not fulfil either of our conceptual criteria for a regulatory earnings metric, and thus will be a poor indicator of whether a course delivers positive outcomes for its students.

However, targeting (c) would not be completely alien to the current OfS approach but would merely require a change in emphasis. This is because the provider-specific benchmarks for other indicators that are already calculated by the OfS, which currently play a subsidiary role in regulation, can be interpreted as estimates of the counterfactual outcomes where a course's graduates had instead taken an average of courses in the same subject. Adopting (c) as the target for an earnings metric could thus be implemented by emphasising a provider-specific benchmark for earnings in an updated

regulatory framework rather than an absolute threshold. In line with this framing, much of the remainder of this report is concerned with different options for estimating the provider-specific earnings benchmark.

This raises the question of what it is that we would expect to cause some courses to overperform and others to underperform their course-specific earnings benchmark. Especially for non-vocational courses, the link between course quality and graduate earnings can be quite indirect, as knowledge acquired on the course does not necessarily translate into skills that are useful in the labour market. Nevertheless, previous work has found that ‘education inputs’, such as teacher quality, class sizes or contact hours, matter for graduate earnings.⁴ Other possible factors could be differences in curriculum content (e.g. whether general-purpose skills such as writing or programming are taught) or differences in explicit labour-market preparation (e.g. through careers services). As we discuss in detail below, factors that providers cannot directly control, such as provider selectivity or location, likely also play a role.

It is worth noting that even target (c) could conflict with the second conceptual criterion above – that a metric should only capture the influence of factors that providers can control – if there are factors that are not directly related to providers’ efforts yet contribute to a course’s causal effect on earnings. Such factors might include a provider’s location (for example, attending university in London may make it easier for students to access the higher-paid employment opportunities in the City because it is easier for them to attend interviews, because London firms are more present on campus or because it is easier for students to find the accommodation necessary to gain work experience) or a provider’s selectivity (which could directly influence graduate earnings through reputation or peer effects).⁵ If the goal was to isolate the causal effect of providers’ efforts to prepare students for the labour market, we would wish to disregard the influence of these factors beyond providers’ direct influence, whereas they might be quite important for a provider’s impact on graduate earnings.

We explore options below for a regulator to present supplementary information alongside a metric targeting (c). In particular, we recommend that, as in the existing OfS B3 dashboard, actual average graduate earnings on a given course (in pounds) should be presented at the same time. Juxtaposing the earnings metric with actual earnings numbers would give stakeholders a sense of whether, for example, graduates from a course performing poorly relative to other courses had low absolute earnings or merely high benchmark earnings. In Section 8.vi, we outline options for how an ‘adjusted benchmark’ could be produced that would incorporate factors arguably beyond providers’

⁴ For instance, Arteaga (2018) studied a 2006 curriculum reform that reduced contact hours by 20% and 14% in economics and business, respectively, at a top university in Colombia. She found that this led to declines in wages for graduates of these courses of 16% and 13%, with the declines seemingly attributable to students’ lower success rates in employer recruitment rounds.

⁵ By selectivity, we mean the prior attainment required for admission to a given provider. A provider’s selectivity will be closely related to the average academic ability of its intake.

control, such as a provider's location or its selectivity. This 'adjusted benchmark' could also be presented alongside our other estimates.

Practical criteria

In practice, there are data and other practical constraints that limit how close an earnings metric can get to any conceptual ideal. Different options will deviate from the target in different ways. We therefore need practical criteria to judge different feasible earnings metrics against each other.

This section lists six high-level practical criteria for an earnings metric. The extent to which these criteria are achieved by our baseline metric is discussed in Section 7.

i. Fairness across providers

The higher education sector is extremely diverse, covering a huge range of subjects and catering to a vast range of students with different characteristics, skills and interests. Different courses differ radically in the material they teach and the kinds of students that enrol on them. A regulatory earnings metric needs to cut through this complexity and compare courses' earnings outcomes on a common scale.

As set out in the section on conceptual criteria above, the key task in making different courses comparable is approximating the causal impact of a course on earnings. While in practice it is impossible to know what students on a given course would have earned had they taken a different course, this can be approximated by comparing the outcomes of students with similar observed characteristics such as prior attainment at school or socio-demographic characteristics.

Whether this is a good approximation to the causal effect will depend on whether there are enough students with similar observable characteristics taking different courses and on how much students differ in their unobserved characteristics that affect their future earnings, such as 'grit' or their access to networks in the labour market. To the extent that students on different courses differ in unobserved characteristics, the resulting earnings metric will be systematically biased against providers taking on students with unobserved characteristics that cause them to earn less in their future careers.

The size of this bias is an important practical criterion for judging an earnings metric. There is no direct measure of its size, as the true causal effect is unknown. However, for any given set of observed characteristics, it is reasonable to conclude that this set is insufficient for minimising the extent of bias if the inclusion of additional control variables substantially changes the estimated coefficients.⁶ Similarly, it would be implausible that

⁶ Variations of this idea have been formalised in the econometrics literature, with practical solutions for producing bounded estimates that allow for the effects of unobservable selection (e.g. Oster, 2019). However, we do not believe that implementing these methods would be practical in this context.

the bias of estimates from any particular model was minimal if seemingly small changes to the modelling approach dramatically changed the results.

Importantly, unobservable characteristics are much more likely to be similar across courses in the same subject area. For example, graduates from two different business studies courses are much more likely to have broadly similar career aspirations than, say, a business studies graduate and a creative arts graduate. This means that there is also a strong practical argument for favouring an earnings metric targeting returns relative to other courses *within a given subject area*.

But whatever the design of an earnings metric, it will not be possible to eliminate bias and thus unfairness across providers entirely. For example, a course in development economics will likely attract students with different career aspirations from a typical economics course, leading to a bias against providers offering such courses. However, this by itself is not a sufficiently compelling reason not to use an earnings metric in regulation. The OfS will be able to view an earnings metric within the context of its existing suite of regulatory metrics and take contextual information about providers and their students into account alongside them. Indeed, this is an important part of the OfS's existing regulatory approach, which considers whether contextual information submitted by the provider justifies comparatively low performance.

ii. Accuracy

A second important practical criterion is the accuracy of an earnings metric. Even if a metric had little or no bias, it would still only be useful if the estimate for any provider was close to the true target with a reasonable probability.

Again, because the true value for any target of an earnings metric is unknown, it is not possible to say how accurate a given estimate is. However, unlike for bias, we can calculate the standard deviation of an estimate arising from statistical noise, albeit only under the assumption that a given statistical model is correct. A more informal but perhaps more robust approach is to compare estimates across years: if the estimate for the same course varies implausibly widely across years, it seems reasonable to conclude that the earnings metric does not accurately measure the targeted causal effect.

iii. Positive incentives for providers

A well-designed metric would generate incentives for higher education providers (HEPs) to improve the labour market outcomes of their graduates. In contrast, a poorly designed metric could create perverse incentives for providers. Of particular concern here are selection incentives. If providers knew that admitting certain types of students would in expectation lead to a lower earnings metric, this could harm university access for such students. This is a particular concern for disadvantaged students and students with protected characteristics.

While it is, in any case, desirable to adjust as far as possible for the characteristics of student intake to minimise bias (see the first criterion above), it is particularly important to take into account characteristics relating to disadvantage to avoid distorting providers' admission incentives. A good measure would adjust fairly for differences in students' existing characteristics in a way that made providers indifferent about selecting them.

Policymakers may also wish to avoid metrics that create perverse incentives for providers to offer particular types of provision, especially where this would damage skills pipelines. For example, the country needs to train nurses, who will likely get a smaller earnings boost from their degrees than investment bankers, but a labour-market outcomes metric should not create incentives for HEPs to stop training nurses and only train investment bankers. This is more of a concern for earnings metrics targeting returns relative to other subjects or relative to non-graduates.

iv. Regulatory practicality

The metric should be designed to fit within the current regulatory framework used by the OfS. Amongst other things, the OfS regulates HEPs on a range of 'B' conditions, which are to do with quality, standards and outcomes.

As argued in Section 2, this metric would naturally fit within the metrics assessing providers on condition 'B3', which cover continuation, completion and progression. Progression is currently measured based on the share of graduates who are in 'managerial or professional employment', doing further study, travelling, caring for someone or retired 15 months after graduation. The new labour-market outcomes metric would most obviously complement or replace this progression metric.

In its regulation of the sector, the OfS focuses on absolute *sector*-level 'thresholds'. However, it also estimates *provider*-level 'benchmarks', which can be further broken down by subject and various other 'splits'. For example, for the progression metric, the threshold for full-time first degree undergraduates is 60% across the sector, but the benchmark is 85.1% for Imperial College London and 70.2% at Middlesex University (it is higher at Imperial College largely due to the high average prior attainment of Imperial students).

As discussed under 'Conceptual criteria' above, an earnings metric satisfying our conceptual criteria can be interpreted as the (percentage) difference between actual earnings and a benchmark earnings level. It would therefore be advantageous for the benchmark earnings level underlying the earnings metric to align closely with the existing benchmarks for other indicators produced by the OfS. One important aspect of this would be to align with the current OfS practice of taking subject of study into account in the estimation of the benchmark.

The second important consideration related to regulatory practicality is that the OfS has its own way of defining subjects, 'levels' and 'modes' of study. For subject, it uses the

'Common Aggregation Hierarchy' or CAH definitions (specifically, it uses 'CAH2'). For level, it includes 'first degree undergraduate' but also other undergraduate degrees, postgraduate degrees and apprenticeships. Furthermore, these levels are split by 'mode' of study, meaning part-time or full-time. It will not be possible for all levels and modes to be included in our analysis: some of these routes do not have many students; we have no data on apprenticeships; and we have agreed with the DfE that postgraduate degrees are beyond the scope of this work. But where possible, we will align the construction of the metric with existing definitions used by the OfS.

v. Timeliness

It is also important from a regulatory perspective for an earnings metric to be tied as closely as possible to providers' performance in the present rather than in the past. This creates a trade-off between the conceptual goal of capturing a course's impact on graduates' earnings over their whole lifetime, and the practical regulatory goal of having an indicator of providers' performance that is as up to date as possible. This trade-off is quite stark: for instance, earnings outcomes in their 40s are likely a key part of the absolute earnings return to their courses for most graduates; yet an earnings metric based on such earnings would be of little regulatory value today, as the youngest students for whom data on earnings in their 40s are available mostly attended university in the 1990s.

Hence there is a further practical argument for favouring an earnings metric based on comparisons within graduates of a given subject. Graduates of courses in the same subject area are likely to have more similar earnings trajectories, as they have access to a more similar set of career options. As a result, earnings differences between graduates of different courses five years after graduation are more likely to be a reasonable guide to earnings differences ten or twenty years after graduation within the same subject than across subjects or compared with non-graduates.

vi. Transparency

For any metric to be perceived as a legitimate performance indicator, it is important that non-specialists can understand the basic logic behind its design. A sophisticated approach may have the benefit of dealing with a concerning property of the metric, but at the cost of preventing a non-specialist audience from engaging with the results. In particular, a policymaker may wish to ensure that university administrators can interpret the results, have confidence that they are 'fair', and appreciate how changes to university practices that affected graduates' outcomes would be reflected in future results.

Transparency and simplicity are particularly important for any earnings metric that relies on confidential data such as the DfE version of the LEO dataset (described in Section 4), as external researchers will not be able to directly replicate any findings. Furthermore, using a well-established methodology with low computational requirements would keep

down the cost for the OfS of generating new estimates as additional years of data became available. Conversely, the cost of complicated estimation methods that would be difficult to extend to future cohorts, that would require large computational power or that would require specialist software could be large.

4. Data

Reliable measures of graduate outcomes need to be based on reliable data. Existing graduate outcome measures used for regulatory purposes by the Office for Students are based on data from the Graduate Outcomes survey which, as discussed in Section 2, captures less than half of all graduates. Relying on the Graduate Outcomes survey for earnings data would be especially problematic, as even those who do respond to the survey may supply inaccurate or imprecise earnings information. In addition, the Graduate Outcomes survey is conducted 15 months after graduation, before many graduates have ‘found their feet’ in the labour market.⁷

Given these issues, we recommend instead making use of administrative tax records for a regulatory earnings metric. Tax records have been linked together with higher education records to give detail on the annual taxable earnings from employment for more than 95% of UK higher education graduates. The resulting Longitudinal Education Outcomes (LEO) dataset allows us to obtain accurate earnings information later in graduates’ careers.

The LEO dataset links together:

- primary school, secondary school and school sixth-form records, as well as GCSE and A level exam results, from the National Pupil Database (NPD);
- higher and further education records from the Higher Education Statistics Agency (HESA)⁸ and the Individual Learner Records (ILR); and
- His Majesty’s Revenue and Customs (HMRC) data, including annual taxable earnings from employment (PAYE) and self-employment and partnerships (from self-assessment tax returns), as well as employment outcomes (whether in sustained employment, in receipt of benefits etc.).

The dataset that has been made available to us – and the one that is likely to be made available to a regulator – is **organised by graduation cohort**.⁹ **This means that our analysis will exclude higher education dropouts.**¹⁰

Following the OfS’s regulatory practice, we split the data by mode of study (full-time or part-time) and level of study. As the scope of our work is restricted to undergraduates, we only have three levels of study: first degree, other undergraduate (Other UG) and

⁷ However, note that the situation is much improved relative to the previous Destinations of Leavers from Higher Education (DLHE) survey, which was conducted approximately 6 months after graduation.

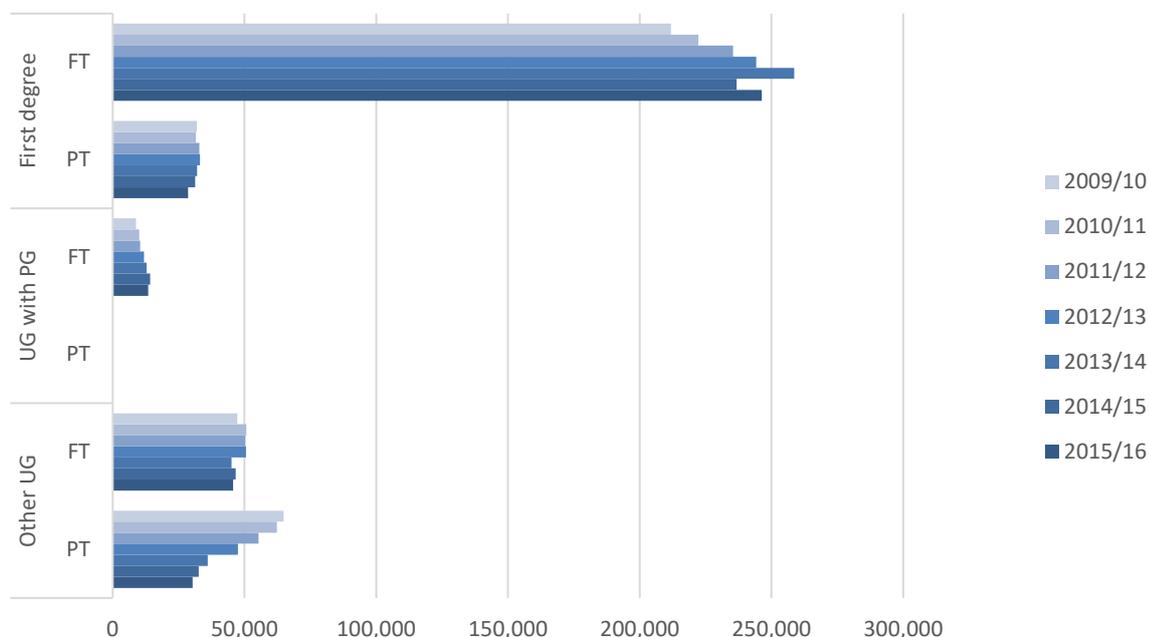
⁸ HESA data include data on Alternative Providers.

⁹ This is different from the dataset that is typically made available to researchers through the Office for National Statistics’ Secure Research Service.

¹⁰ A concern with this is that it could in theory create an incentive for HEPs to fail students that they did not expect to do well in the labour market. However, we do not expect this to be an important issue in practice, as HEPs are assessed separately on their completion rates by the regulator and these also feed into university league tables.

undergraduate with postgraduate components (UG with PG). Figure 1 shows the number of graduates by level, mode and academic year of graduation in the data we will use. As part-time UG with PG courses are very rare, we focus on the other five level–mode combinations, of which full-time first degree is by far the largest category.¹¹

Figure 1: Number of graduates by level, mode and academic year of graduation



Note: Includes all England-domiciled graduates in HESA or ILR data from undergraduate courses, whether or not they can be matched to school or tax records.

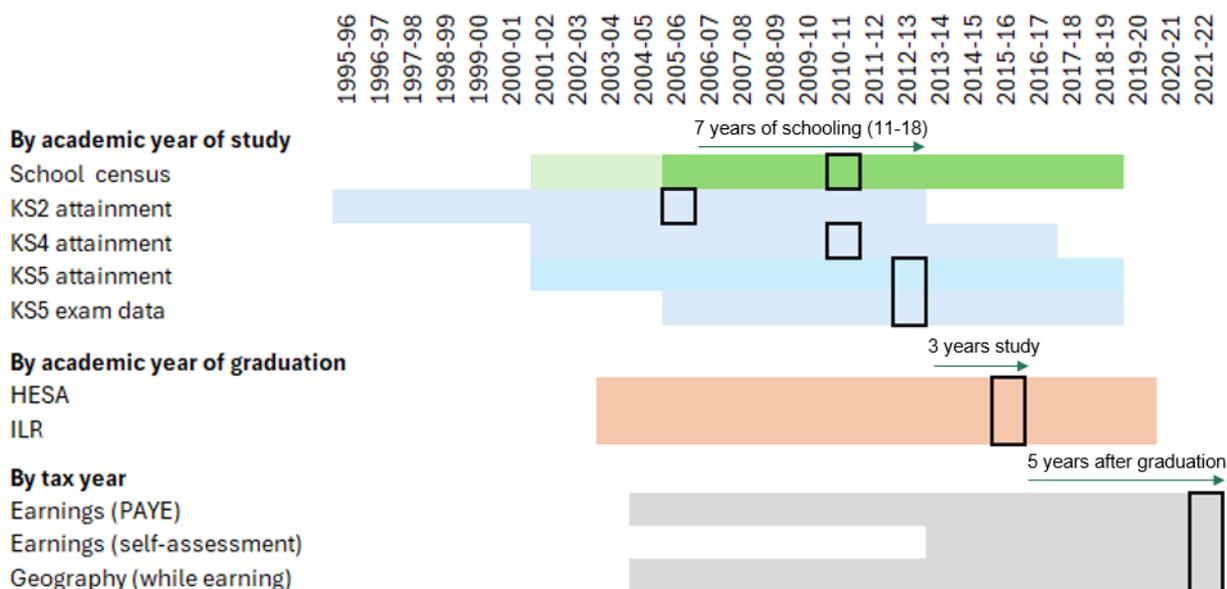
The 2009/10 to 2015/16 graduation cohorts shown in Figure 1 are the main graduation cohorts we will use in the analysis for this report. As we discuss below, our main analysis will focus on earnings up to five years after graduation, meaning later graduation cohorts are not usable, as full earnings data are not yet available for them.¹² Earlier cohorts are not usable because GCSE results are only available from the 2001/02 GCSE cohort onwards, so we would not observe GCSE results for too large a share of graduates from earlier graduation cohorts. We would also not be able to observe all self-employment earnings from earlier cohorts, as self-assessment data are only available from the 2013/14 tax year.

¹¹ Our categorisation into levels is slightly different from the categorisation used by the Office for Students. The OfS categorises all veterinary, dental and medical graduates who have obtained eligibility to register to practise with their respective professional bodies as obtaining UG with PG degrees. We cannot do this, as we do not observe eligibility to register with professional bodies in our data. We therefore classify these courses as first degrees.

¹² In general, as self-assessment tax returns do not have to be filed until January after a tax year has ended, there will be at least a one-year lag between the end of a tax year and self-assessment data being available. We judge that the advantages of including self-assessment tax data in the earnings metric would outweigh the disadvantage of the additional delay relative to PAYE data.

Figure 2 visualises which years of data we have from which data sources, and how that relates to a typical path through the education system. The highlighted fields show the main data items required for the 2015/16 graduation cohort for our baseline model. This cohort enables us to consider earnings up to five tax years after graduation, as the latest year of earnings data we have available is 2021/22. Those graduating from a typical three-year degree will have started their course in 2013/14.

Figure 2: Years of available data from different administrative data sources



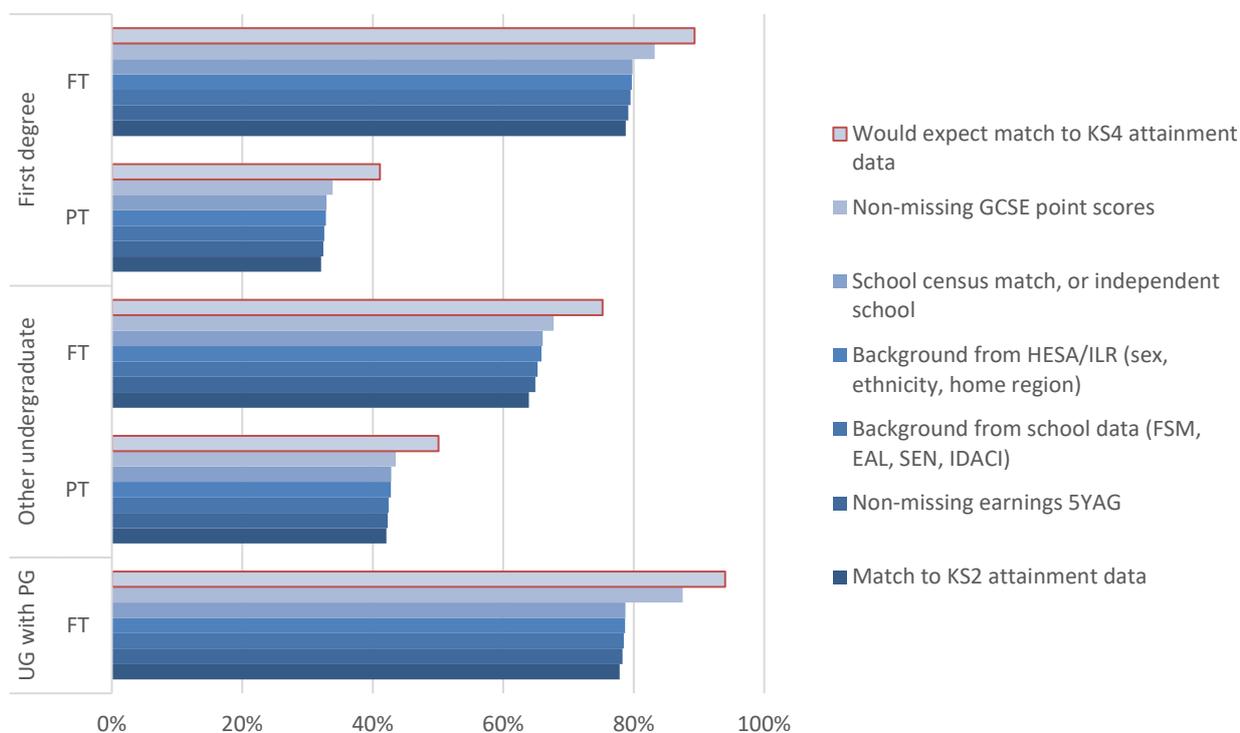
Note: Coloured fields indicate available data. Years are academic years for education data and tax years for tax data. Highlighted fields show the main data items required for the 2015/16 graduation cohort for our baseline model.

Those entering university straight after their school-leaving exams sat their A levels (or equivalent) in 2012/13, their GCSEs in 2010/11 and their Key Stage 2 exams (SATs) in 2005/06. But some degrees take longer than three years, and many students do not enter university straight after leaving school. As a result, a large fraction of the 2015/16 graduation cohort will have sat their GCSEs before 2010/11, and a small share will have sat them before 2001/02, when GCSE records first become available.

As shown in Figure 3, not observing school records for mature students is the biggest threat to coverage for our data. This is not a major issue for full-time first degree or UG with PG students, where 89% and 94% of graduates respectively were young enough at entry that we would expect them to match to the school records. However, for some other levels of study the percentage is much lower: only around 40% of part-time first degree students, 75% of full-time other undergraduates and 50% of part-time other undergraduates are matchable to the school records. Estimates for these groups will therefore be based on the relatively younger students who completed these courses (as it is the younger students who we are able to match to the school records). This could generate biases if the younger students are not representative of all students on the

course, but perhaps more pertinently it will harm our ability to produce reliable estimates in some of our models.

Figure 3: Match rates between data sets for the 2015/16 graduation cohort



Note: 'Would expect match to KS4 attainment data' is based on the presence of the anonymised identifier (Pupil Matching Reference) in the data which is required for a match to school records, and is largely explained by age at entry into higher education. 'Background from school data' is only required for those not recorded in the KS4 attainment data as attending an independent school in Year 11.

The issue of mature students aside, match rates with the administrative data are extremely good. Of those starting their courses at a young enough age that we would expect them to be in the school data, we can match the vast majority – more than 92% – to their actual GCSE attainment data. Of these, nearly all have non-missing GCSE point scores, can be matched to the school census (or went to independent school), have university and school record information on relevant background characteristics, and can be matched to an HMRC earnings record.¹³ A large majority can also be matched to Key Stage 2 attainment data (later, we will use the lower match rate as justification for excluding the Key Stage 2 data from our baseline specification).

It is worth noting that the shares of students we can match to school records will improve over time as the graduation year moves further and further away from the first year for

¹³ The main source of missing GCSE point scores is people who took English or Maths GCSEs a year early, which is the norm at some independent schools. It is in principle possible to obtain these scores from the previous year's exam records, but this cannot be done with the data we have been given. All of our estimates therefore exclude graduates who took the English or Maths GCSEs a year early. This is unlikely to affect our estimates appreciably outside of a small group of mostly highly selective providers for which independent school graduates make up a large share of the student body.

which GCSE results are available (2001/02). This means that the reliability of any earnings metric will similarly improve over time. However, given the length of part-time courses and the prevalence of older students taking them, for part-time courses it will still take several years for GCSE results to be available for anywhere near the full cohort.

Our data do not currently include students whose school exams were disrupted by the COVID-19 pandemic in 2019/20 and 2020/21. For our baseline specification, this will not become relevant for some time, as we do not make use of A level data and, leaving aside highly unusual cases, it will take until 2031 before the first graduates from these cohorts will have completed five tax years after graduating from their undergraduate degrees.

In our view, it would be reasonable for a regulator to simply replace exam grades with the teacher-assessed grades in those years, as we standardise all GCSE scores within cohort. While any bias in teacher-assessed grades may feed through to bias in our earnings metric estimates, we would expect the resulting bias to be minor.¹⁴

¹⁴ An alternative would be to interact some or all control variables with a dummy for having teacher-assessed GCSE scores. This could lower bias but would likely come at some cost in terms of efficiency.

5. Key analytical decisions and introducing our ‘baseline’ approach

There is a large set of decisions that need to be taken in the formulation of an earnings metric. In this section, we classify those decisions into five groups and discuss some of the options within each of these groups with reference to the conceptual and practical criteria outlined in Section 3. We also document the decisions we have taken for what we are calling our ‘baseline’ approach. In later sections, we will test the sensitivity of our estimates to these decisions.

i. Counterfactual

As discussed in Section 3, a key decision in formulating an earnings metric is the counterfactual against which graduates’ actual earnings are compared. Previous research in this area has focused on the average returns to attending a particular HEP, doing a particular subject or doing a particular course *relative to not attending university at all* (e.g. Belfield et al., 2018; Britton et al., 2020) or *relative to a baseline case within the higher education sector* (e.g. Britton et al. (2022), who use history at Sheffield Hallam University as a base case).

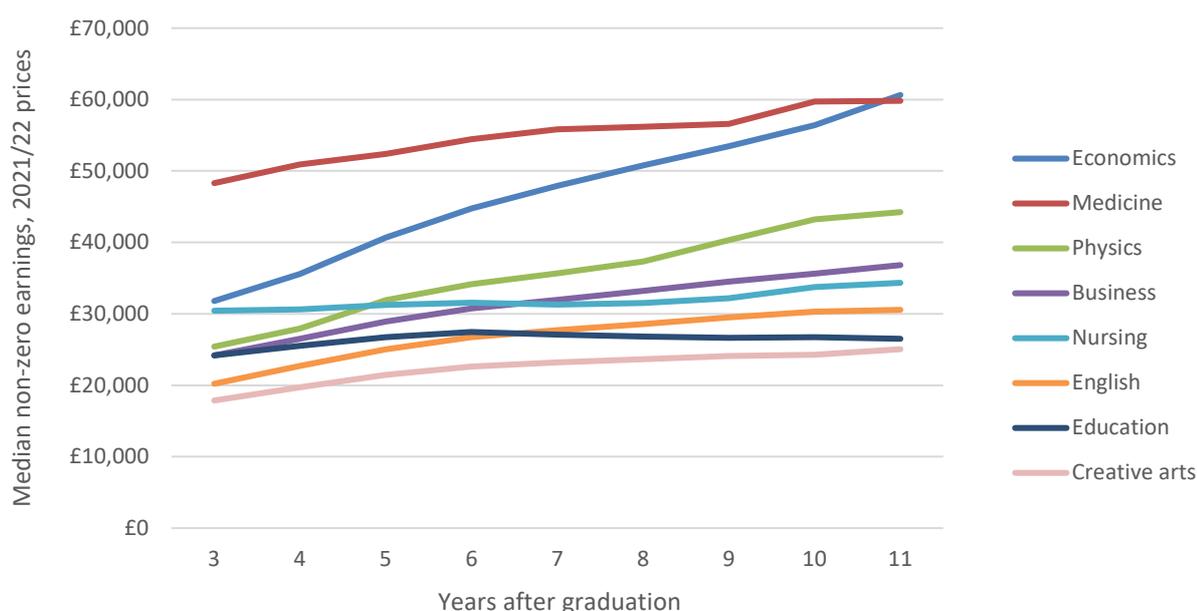
As we have argued in Section 3, there are strong conceptual and practical reasons to target the causal effect of university courses *relative to an average of courses within each subject area*. We briefly recap the main arguments here:

- **Conceptually, the causal effect relative to an average within a subject area is most closely related to provider effort.** Absolute earnings returns also depend on factors influencing counterfactual non-graduate earnings such as minimum wage legislation or vocational training opportunities in students’ local areas. Relative earnings returns relative to an overall average of university courses also depend on the labour market value of skills in a particular subject area, which is determined by the demand for and supply of such skills in the wider economy rather than by any particular provider’s performance.
- **Unobserved characteristics are much more likely to be similar across courses in the same subject area.** For example, graduates from business studies courses are much more likely to have broadly similar career aspirations than, say, business studies and creative arts graduates. As a result, estimates are subject to much less bias, making the earnings metric fairer.
- **Assessing courses’ earnings effects within each subject area would avoid creating incentives for providers to concentrate provision in subjects with the largest positive impact on pay.** While creating such incentives could be an intended effect for policymakers, doing so would not come without risks, including possible damage to skills pipelines for the state education sector, the

NHS and the creative industries. It is also not obvious that university regulation – rather than, for example, differential teaching grants as proposed by the Augar Review¹⁵ – would be the most appropriate tool for creating these incentives.

- **The current OfS provider benchmarks take subject of study into account.** This means that an earnings metric targeting the causal effect of university courses relative to an average of courses within each subject area would fit in more naturally with the existing measures produced by the OfS.
- **Graduates of courses in the same subject area are likely to have more similar earnings trajectories, as they have access to a more similar set of career options.** As a result, earnings differences between graduates of different courses five years after graduation are more likely to be a reasonable guide to earnings differences ten or twenty years after graduation within the same subject rather than across subjects or compared with non-graduates.

Figure 4: Median earnings by subject area and years after graduation



Note: Median annual taxable earnings each full tax year after graduation for the cohort graduating from first degrees in the 2009/10 academic year and who studied full-time. This covers tax years 2013/14 to 2021/22 and includes earnings from PAYE and self-assessment. Results shown for selected subject areas (CAH2), with individuals on joint courses weighted by full-person equivalents. Includes all those with positive earnings in a given tax year and is not restricted to those with school attainment data. Figures are CPI real, in 2021/22 prices.

¹⁵ <https://www.gov.uk/government/publications/post-18-review-of-education-and-funding-independent-panel-report>.

The last of these points is the only one for which we can present reasonably direct empirical evidence. As shown in Figure 4, the differences in early-career earnings trajectories across different subject areas are substantial. Graduates from degrees that are more vocational, such as education, nursing or medicine, experience high earnings early in their careers. Graduates from subjects such as physics or business are slower to get going but tend to do very well in the medium and long term.¹⁶ These differences militate in favour of comparing courses in the same subject area.

For all these reasons, our **baseline approach will be to estimate a ‘relative’ measure of a course’s earnings return, within a given subject area.** This indicator will be of interest to a regulator seeking to identify institutions that are falling behind their ‘competitors’ in the same subject area with regard to their graduates’ earnings.

It is important to be clear that this ‘relative’ measure will in some cases give very different results from the ‘absolute’ earnings return delivered by a course. In a subject such as law or economics with a high absolute return, a given course might score poorly relative to other courses in the same subject area even if it increases students’ earnings relative to not going to university. Conversely, in a subject area with low absolute returns, such as creative arts, a course might score well on this relative measure even if it fails to increase students’ earnings relative to not going to university.

ii. Outcomes

The tax data do not give us a wide range of options to consider for the outcome variable, as they do not include much more information than the amount an individual earned in a given tax year (e.g. they do not include information on the hours they worked or the job they were doing).¹⁷ We are therefore limited to small variations on the following options:

- earnings;
- the probability of earning above a given earnings threshold.

Our main focus will be on the former outcome variable, although we will also discuss options for the latter in less detail. **Specifically, we use the natural logarithm of total earnings as our outcome variable, as is standard in economics.** Reasons for taking the logarithm of earnings include the fact that empirical earnings distributions are roughly log-normal, and the common finding that returns to education tend to be multiplicative (i.e. similar in percentage terms across individuals) rather than additive (i.e. similar in pound terms across individuals).

We use total earnings in a reference tax year – i.e. the sum of employment earnings, self-employment earnings and partnership profits, excluding any losses – as this is the

¹⁶ Gender composition differences across subjects substantially affect these trends, but the point stands. Gender-specific versions of Figure 4 are shown as Figures 23 and 24 in the appendix.

¹⁷ In addition to earnings, we have information on the industry of employment, but we have not used it for this report.

most comprehensive measure of earnings available.¹⁸ However, one might still be concerned that an earnings metric based on this outcome variable might disadvantage courses whose graduates are especially likely to set up their own businesses, as these ventures may take time to deliver substantial profits. Reassuringly, as shown in Figure 39 in the appendix, our baseline earnings metric is very weakly *positively* correlated with the share of graduates with self-assessment earnings, making it seem unlikely that such courses are systematically disadvantaged by our baseline earnings metric.

Another important concern with using total earnings in a tax year as the outcome variable is that, especially for those with very low earnings, these earnings might not be representative of their true earnings potential. For instance, low earnings might be due to someone working part-time or being out of the labour force for part of the tax year (e.g. due to parental leave). To guard against the bias in our estimates that could be introduced by such cases, it makes sense to exclude from our estimates graduates for whom we observe no or very low earnings. **Our baseline approach excludes all earnings observations of less than £3,000 (in 2021/22 prices).**

At the other end of the earnings distribution, a concern is that outliers with extremely high earnings, such as popstars or professional football players, could bias estimates for the courses from which they graduated. In our baseline approach, we **Winsorise earnings at the 99th percentile within bucket and sex**. Winsorising means adjusting earnings above a certain level (here the 99th percentile within bucket and sex) down to the earnings at that level (the same procedure is known as ‘top-coding’ when it is done for data privacy purposes). This ensures that outliers are removed while preserving sample size and the shape of the earnings distribution for different subgroups.¹⁹

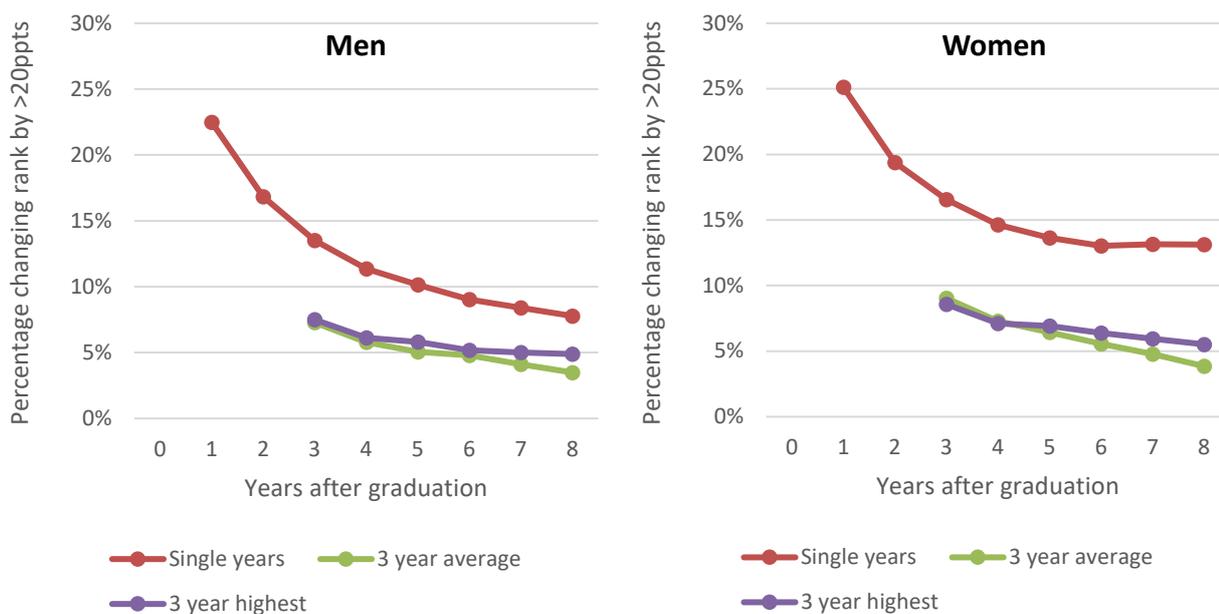
A key complementary decision is at what point after graduation these outcomes should be measured, and whether to use one or multiple years of earnings data. As discussed above, there is a stark trade-off between timeliness, which is crucial for regulatory purposes, and capturing the true causal effect of a course over a graduate’s whole lifetime. Pooling multiple years of data would therefore be advisable if and only if the result was a better reflection of a graduate’s (relative) lifetime earnings. This would depend on the properties of the earnings process: for instance, pooling could help if graduate earnings followed a deterministic trend with transitory shocks but would be counterproductive if earlier earnings were uninformative about later-life earnings.

While a formal investigation of the statistical properties of the graduate earnings process is beyond the scope of this report, we do show some empirical evidence here to motivate our choices for the baseline measure. Figure 5 considers a measure of volatility in early-career earnings: the share of individuals who jump up or down in the graduate earnings distribution by more than 20 percentiles from one year to the next.

¹⁸ For the avoidance of doubt, our total earnings variable is the sum of PAYE and self-assessment earnings for the same tax year, even though self-assessment tax returns are only submitted to HMRC in the tax year after the tax year to which they relate.

¹⁹ In contrast, trimming would exclude the observations above a given earnings level from the sample.

Figure 5: Share of graduates changing earnings rank by more than 20 percentiles



Note: All data are for full-time first degree students who graduated in academic year 2012/13.

The red line shows that this is around a quarter of graduates at the very start of their careers, dropping to around 8% for men and 13% for women eight years after graduation. This suggests that early-career earnings are subject to considerable shocks, but that earnings ranks within a graduation cohort have largely stabilised five years after graduation. By that time – likely mostly because of earnings losses associated with childbearing – women are much more likely to experience large changes in their earnings rank than men.

The green and purple lines show alternative measures that pool over three years: the green line takes the moving three-year average, while the purple takes the highest earnings recorded in the three-year period. Both these measures are (by construction) much more stable than earnings in any given year, with fewer individuals changing earnings rank substantially by these measures between years. As we are especially concerned about the distorting effects of temporarily lower earnings (e.g. due to parental leave), our **baseline approach will take the highest earnings recorded in the third, fourth and fifth tax years after graduation.**

iii. Methodology

The target of the methodological approach is to estimate the causal effect of attending the HEP relative to other HEPs in the same level–mode and subject area. Our preferred **baseline approach is to estimate an Ordinary Least Squares regression model, controlling for student characteristics, with HEP fixed effects.**

Technically, this can be written as

$$y_{ib} = \alpha_{0b} + \sum_{j=1}^{J-1} \gamma_{jb} h_{ij} + \sum_{k=1}^K \rho_{kb} x_{ik} + \epsilon_{ib} \quad (1)$$

where y_{ib} is the outcome variable for individual i in bucket b (where bucket is the combination of subject and level–mode). As described in the previous subsection, we will use the natural log of the highest earnings recorded in the third, fourth and fifth tax years after graduation in our baseline approach. The dummies $h_{i1}, \dots, h_{i,J-1}$ are the set of dummies for all institutions within the bucket, with one omitted HEP. The variables x_{i1}, \dots, x_{iK} are a set of student-level controls for background and prior attainment (see more detail on this in the next subsection, iv). These controls could be any combination of categorical and continuous variables.

In this example, the γ coefficients approximate the causal impact of attending the associated HEP on early-career earnings *relative* to the omitted HEP within the bucket. For example, in Britton et al. (2022), the omitted category was Sheffield Hallam University, and so all estimates were the percentage return to attending a given university *relative to attending Sheffield Hallam University*. However, this is not practical for regulation, as ideally the regulator would have estimates for *all* HEPs in any bucket (so long as they are sufficiently large). We will therefore adjust equation (1) by excluding the constant and de-meaning all the x variables and the outcome variable. This means that we get an estimate for each HEP included in the model, and that the interpretation of this coefficient estimate is then the impact of attending that HEP on earnings *relative to the average effect* (weighted by the number of students) of attending any HEP in the sample. Technically, these are the β coefficients²⁰ from the model

$$\tilde{y}_{ib} = \sum_{j=1}^J \beta_{jb} h_{ij} + \sum_{k=1}^K \delta_{kb} \tilde{x}_{ik} + \eta_{ib} \quad (2)$$

where \tilde{y}_{ib} is the de-meaned outcome variable and $\tilde{x}_{i1}, \dots, \tilde{x}_{iK}$ are the de-meaned x variables.

From this model, we can construct a ‘benchmark’ level of earnings for each HEP, which can be thought of as the ‘expected’ level of earnings for a given course, given the composition of its graduates. For HEP j , we calculate this by predicting the outcome variable for each student at HEP j , without including the estimated effect, such that the expected outcome variable for individual i , denoted \hat{y}_{ib} , is equal to a weighted sum of their background characteristics, where the weights, denoted $\hat{\delta}_{1b}, \dots, \hat{\delta}_{Kb}$, are the estimated coefficients from equation (2):

²⁰ Because the outcome variable in our main specification is in natural logarithms, for our final estimates we convert the β s into percentage points, via the formula $100(\exp(\beta) - 1)$. We also adjust our standard errors accordingly.

$$\hat{y}_{ib} = \sum_{k=1}^K \hat{\delta}_{kb} \tilde{x}_{ik}$$

We then take the average of this predicted value across all students at HEP j that are included in the estimation of equation (2). Since our baseline model is in natural logarithms, in a final step we exponentiate this average predicted value to get a cash-terms benchmark for the provider. We would expect, for example, that for graduates of most subjects, all else equal, a student's maths GCSE score would be a strong predictor of earnings, resulting in a large and positive coefficient estimate on maths GCSE scores in equation (2). All else being equal, this would then translate to higher benchmark earnings for a provider that admits students with very high maths GCSE scores than for a provider that admits mostly students with low maths GCSE scores.

We obtain heteroskedasticity-robust standard errors for the β s using the standard degrees-of-freedom adjustment ('HC1').²¹ These standard errors are biased in small samples, but valid asymptotically even where different observations have different error variance (heteroskedasticity).²²

Some alternative approaches

The fixed effects approach is the one that is most used in the academic literature on this topic.²³ However, there are a few alternative approaches that we want to pay closer attention to later, which we document below.

- **A mean-residual approach.** This is an approach that has commonly been used in value-added models for schools and for teachers. It is a two-step procedure, where the first step is to estimate a model as in equation (2) above but *excluding the HEP dummies*. The second step involves calculating all of the residuals for individuals at each HEP and averaging them. This approach relies on the same 'unrelated effects' assumption as a **random effects** model, namely that the effect for each course is uncorrelated with all control variables (our estimates are so similar to the random effects approach that we do not report both).²⁴

²¹ Confidence intervals are converted to percentage terms according to the formula given in the previous footnote.

²² For the two-stage 'mean-residual' and 'hybrid' estimators discussed below, reported standard errors only take into account statistical uncertainty at the second stage.

²³ It is common in the academic economics literature to combine the fixed effects approach with additional controls for the set of higher education courses each student applied for, in an attempt to control more completely for student preferences (e.g. Dale and Krueger, 2002; Mountjoy and Hickman, 2021). Such an approach would be possible in this setting, but we were unable to access the necessary data to consider it for this report.

²⁴ The terminology here reflects standard usage in economics and econometrics. In other disciplines, the 'random effects' model is known as a 'mixed effects', 'multilevel' or 'hierarchical' model with a random intercept.

A key advantage of this approach is that it would make it possible to control for characteristics that do not vary at the provider level in a fixed effects model, such as the selectivity or geographical location of the provider. In the baseline fixed effects model, this could not be done, because any provider-level controls would be perfectly collinear with the HEP dummies. The main motivation for including provider-level controls would be a desire to abstract from provider-level factors such as selectivity or location which will affect earnings but are not things providers can easily change (as discussed under 'Conceptual criteria' in Section 3).

Note, however, that this approach depends upon a strong assumption, namely that the effects of factors that providers can control to influence graduate outcomes are uncorrelated with both the individual-level control variables *and* any provider-level characteristics included in the model. For example, consider two factors:

- the quality of career services offered by a provider (something the provider can control and that is likely to influence graduates' earnings);
- the selectivity of the provider (a provider-level characteristic that the provider arguably cannot directly control).

The assumption requires that there be no systematic relationship between the quality of career services and the provider's selectivity. But suppose, for example, that more selective universities have better-funded career services. In this case, the model would incorrectly attribute part of the positive effect of high-quality career services to the provider's selectivity, so the provider's actual impact on graduate outcomes would be underestimated.

Similar issues could arise with individual-level controls. If providers with higher-quality career services also had more effective 'widening participation' policies and therefore more graduates that were eligible for free school meals, a mean-residual or random effects approach would underestimate the earnings impact of providers with a high share of free school meal students. The reason is that the model would wrongly misinterpret the effect of high-quality careers services as a less negative effect of socio-economic disadvantage on earnings.

- **A hybrid random / fixed effects approach.** We implement this using another two-step procedure, except that now the *second step* is to estimate the model shown in equation (2). The first step is to *partial out* any of the control variables that only vary at the provider level from the outcome variable and the remaining controls. This is useful because the required assumption is slightly weaker than with the mean-residual and random effects approaches: it still needs to be assumed that the provider-level effects representing factors affecting graduate

earnings that providers can control are uncorrelated with provider-level controls (such as selectivity or geographical region), but they do not need to be uncorrelated with all of the individual-level controls (such as the student's prior attainment or socio-economic background).

- **A non-parametric approach.** This is another approach that is commonly used to estimate value-added models. For example, it is used to estimate Progress 8, the main value-added model used to measure secondary school effectiveness in England. It is also used by the Office for Students to calculate the existing benchmarks for the B3 indicators. The approach is typically implemented by dividing all individuals into 'bins' of people with very similar characteristics, and then comparing each individual's outcome with the average of people within their bin. Someone who is above average in their bin would contribute a positive value-added score to the university they attend, while someone below average for their bin would contribute a negative score. Alternatively, the same result can be obtained using the mean-residual approach outlined above, with dummy variables covering the bins as control variables.

The non-parametric approach works well in the context of Progress 8, where the only background characteristic used is prior test scores, which is a continuous measure that is easy to divide into bins. However, in our context, a broad set of control variables is relevant beyond prior test scores (see more on control variables below). In this case, division of people into bins is more difficult, and it is common to end up with few or no observations in some bins representing graduates with less common combinations of characteristics. In the extreme case of a bin only containing graduates from a single provider, the estimated counterfactual for those students is the same as their actual outcome, so in effect they do not contribute to their provider's benchmark.

This is a serious concern for a regulatory metric, especially given that a provider's performance in supporting students from backgrounds that are less represented might be of particular interest for the regulator. There is a trade-off here: too many bins will leave no comparison group for students with less common characteristics, whereas too few bins will lead to students with substantially different backgrounds being compared with each other. The OfS's current benchmarks are based on a large number of bins (e.g. 22,440 for full-time first degree students); for context, the 'contribution to own benchmark' for each provider is reported (overall and for 'split indicators').

iv. Control variables

As argued in Section 3, it is crucial for an earnings metric to account for the large differences in student composition at different HEPs for both conceptual and practical

reasons. As described above, our proposed baseline methodology is to use a regression model that includes controls for student composition. Here we list the main control variables that we have available, grouped by the dataset that they are available from.

HESA and ILR data:

- Sex
- Broad ethnic group (self-reported)
- Region lived in upon application
- Graduation cohort
- Whether participating in further study

Key Stage 2 (roughly age 11) NPD records:

- Key Stage 2 test scores in English, maths and science

Key Stage 4 (roughly age 16) NPD records:

- GCSE grades in all exams taken, and equivalent qualifications
- Academic year in which student was in Year 11
- School type (including whether independent)
- Free school meal (FSM) eligibility on day of the school census
- Special educational needs (SEN) status (including statemented and non-statemented)
- English as an additional language (EAL)
- Local area deprivation (as measured by the IDACI index), based on student's home postcode

Key Stage 5 (roughly age 18) NPD records:

- A levels and equivalent, grades and subjects

HMRC tax data:

- Home local area as an adult (in addition to the key outcome variables which come from the HMRC data)
- Employment outcomes (whether in sustained employment, in receipt of benefits etc.)

Our baseline approach will *exclude* controls for Key Stage 2 and Key Stage 5 attainment, as well as the controls for where people live as adults. We provide our rationale for this below. **Control variables from all other categories above will be included.** GCSE scores will be converted into standardised scores, which we include as continuous measures, as well as controls for grades in maths and English GCSEs.²⁵ The

²⁵ We do not observe background characteristics from the age 16 school records for the privately educated (FSM eligibility / SEN status / EAL / local deprivation) as privately educated students do not appear in the school census. However, our inclusion of a dummy for being privately educated means that, on the assumption that the earnings effects of socio-economic differences within privately educated students are minor, this should not generate appreciable bias in our coefficient of interest (namely, the HEP effects), though it would likely affect the interpretation of the private dummy.

precise variables included in our baseline approach, and some simple descriptives, are described in Table 12 in the appendix.

Our justification for exclusion of some available controls is as follows:

- **Key Stage 2 records.** As seen earlier in Figure 3, our match rates worsen when we include the KS2 records. Ultimately because including these requires more data and more computational power, we think we should only do so if KS2 scores make any difference to our estimates.
- **Key Stage 5 records.** KS5 qualifications are varied: even amongst A level qualifications, students choose from a diverse range of subjects, with differences in difficulty and in the extent they prepare students for different courses, and different likely effects on future earnings (regardless of students' participation in higher education). Further difficulties arise due to vocational qualifications such as BTECs or T levels. It is therefore not straightforward to control effectively for KS5 attainment. The KS5 records also interact with university admission in a way that demands caution: the final scores are determined *after* offers are made. Controlling for KS5 records therefore generates greater incentives for universities to make lower or even unconditional offers.
- **Region as adults.** This is a controversial choice. A common concern that we have encountered is that an earnings metric might unfairly penalise providers serving the skills needs of their local communities, where earnings are often lower than in strong labour markets (primarily in London). Controlling for region when earning would address this concern, as the resulting earnings metric would only reflect the earnings differences between graduates working in the same regions.

Nevertheless, and in contrast to the OfS's approach to benchmarking, we do not control for region when earning in our baseline approach. The reason is that location choices after university are themselves affected by a student's course, making region when earning a 'bad control'. Controlling for it anyway would move the earnings metric away from measuring the (full) impact of a course on the earnings of its graduates, because an important channel by which providers can have an impact on graduates' earnings would be shut down.

We discuss some alternatives below on how the regulator could address this issue. The simplest option is to contextualise the earnings metric by reporting relevant statistics alongside it, such as the share of graduates staying in the same local area as the provider or the share living in London (see Section 6). In addition, earnings could be adjusted for local living costs (see Section 8.v).

Finally, we also explore directly controlling for provider location (see Section 8.vi).

We will assess the sensitivity of our estimates to the inclusion of controls covering each of these three categories in Section 8.

We have the following variables in the HESA data but not in the specific cut of the ILR data provided to us for use in this project. We do not use these, but future work using a different cut of the data could reconsider this decision.

- NS-SEC classification of parents if under 21 (or student if mature)
- POLAR quintile of area they applied from
- Highest qualification on entry

In its benchmarking for the existing B3 metrics, the OfS also uses care experience, disability, whether local or distance learner, parental higher education and TUNDRA. We have not had access to these variables and so have not investigated whether controlling for them affects our results; again, future work could do this.

Finally, we note that there are alternative approaches to selecting the set of student composition controls that should be included in the model. In particular, machine learning techniques such as LASSO estimators could be used to select the set of control variables (and interactions between them) that add the most predictive power in the model. However, because this is both complex and computationally burdensome, and we see transparency and simplicity as important practical criteria for a regulatory earnings metric (see Section 3), we do not pursue these approaches here.

v. Minimum sample size restrictions and cohort pooling

Our baseline methodology is to estimate separate regression models by level, mode and subject area. Since there are five level–mode combinations within our scope and 34 subjects (aggregating Celtic studies with languages and area studies), this means there is a maximum of 170 level–mode–subject combinations, which we refer to as ‘buckets’. We have to make some decisions about which buckets contain enough students and courses to be included in our models.

First, we need to ensure that there are enough students at a given HEP within a specific bucket for results for that provider to be reported, both because of statistical stability but also to avoid any of the results being disclosive about the earnings any individuals in the data. For example, if there are only three students who did a full-time first degree in mathematics at Balls Pond Road University, we would need to exclude Balls Pond Road University from any reporting of the results for that bucket on the grounds that three is too small a sample for results to be meaningful (given random variation) and because reporting estimates based on such a small number of individuals would be potentially disclosive. The OfS uses a minimum sample size of 23 in reporting statistics to

circumvent these issues, while Belfield et al. (2018) required at least 30 students for a course at a specific provider to be included.

Our baseline approach is to include all providers in our regressions, but to have a minimum sample size of 50 students (with earnings of at least £3,000) at a given provider for us to report estimates for that provider. We believe that a somewhat more conservative approach than in previous work is warranted in this context to ensure the robustness of estimates. Compared with existing OfS indicators, we emphasise the difference between an indicator and its benchmark as our headline metric, whereas the benchmark is only of subsidiary importance for the existing OfS measures. This makes it more important that any model-based estimates are sufficiently robust. Compared with Belfield et al. (2018), we believe that a higher standard of robustness should apply to results for regulatory purposes than to results that are meant to be for information only.

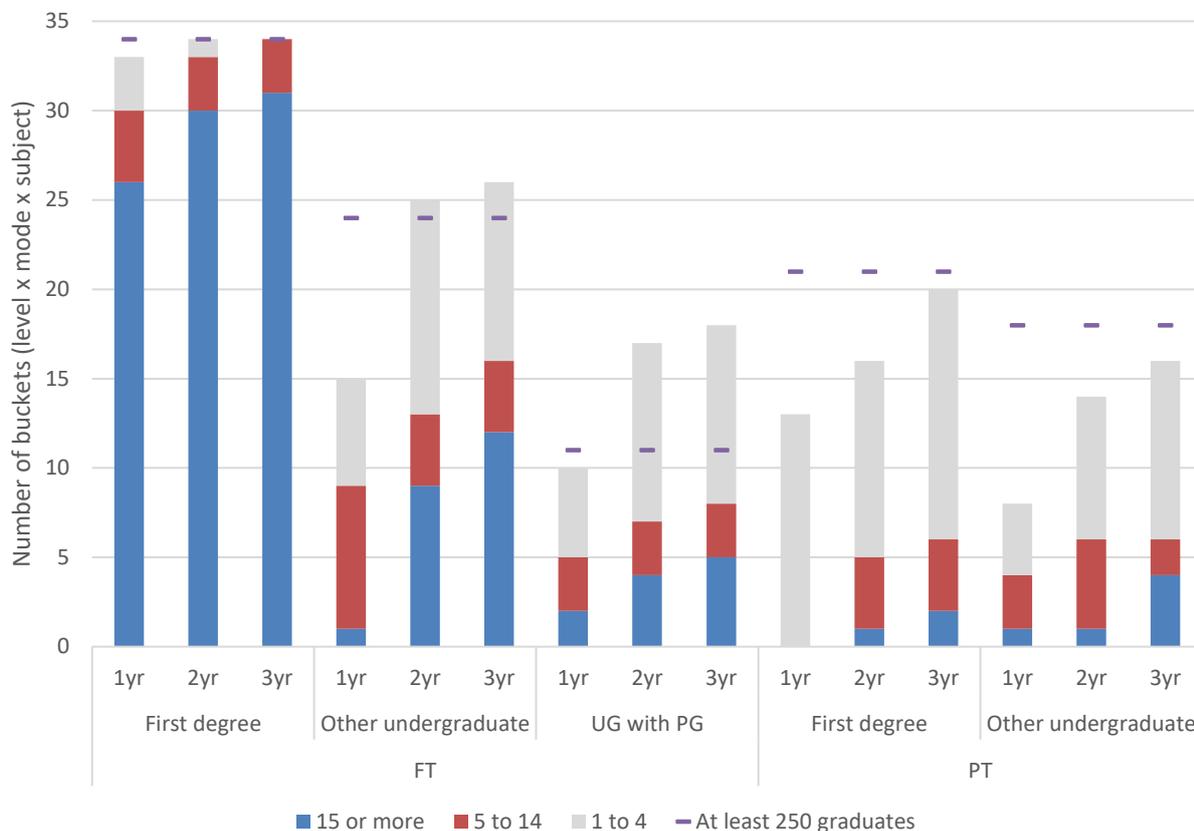
Second, we need to ensure that there are enough providers within a given bucket for any results for that bucket to be included. Our baseline methodology involves estimating a return within a given bucket, relative to other providers in that bucket. Therefore, it is impossible to estimate a model if there is only one provider in a given bucket. We also do not think it is advisable to estimate models if there are only very few providers in a given bucket, as a given provider's results would be unduly influenced by the performance of other providers within the same bucket. **Our baseline approach is therefore to report results only for providers whose students account for no more than 20% of students in a given bucket.** This is similar to a measure used by the OfS of a provider's contribution to its own benchmark.

With these restrictions, we find that many of the 170 buckets are ruled out. We can improve the situation, however, if we pool across cohorts to increase student numbers. **Our baseline approach is to pool across two cohorts.** This choice involves a trade-off: on one hand, increased pooling means more data and hence more estimates, and more stable estimates; but on the other, more pooling means going back further in time. The latest cohort we can look at with earnings up to five years after graduation is currently the 2015/16 cohort: adding one more means including the 2014/15 graduation cohort, and adding more would mean going further back still. As with the decision over the age to look at for earnings outcomes, we must balance the reliability of estimates against going back so far that the estimates are no longer relevant.

Figure 6 shows the number of subjects within each level and mode that we are considering for which different criteria are met, depending on how many years of data are pooled. The height of the bars shows for how many subjects there is at least one provider meeting the minimum sample size threshold of 50 students. Of these, the blue segments show how many subjects have at least 15 providers, the red segments how many have 5–14 providers and the grey segments how many have 1–4 providers. The purple line shows for how many subjects in each level and mode combination there were at least 250 graduates in 2015/16 (which for our baseline restrictions would be the theoretical

minimum of students needed for us to report results for at least five providers for that subject based on just a single year of data). This can be interpreted as the number of subjects within each level–mode combination for which we could plausibly hope to produce estimates.

Figure 6: Buckets for which estimates can be provided by level and mode



Note: The height of the bars shows for how many subjects there is at least one provider meeting the minimum sample size threshold of 50 students. Of these, the blue segments show how many subjects have at least 15 providers, the red segments how many have 5–14 providers and the grey segments how many have 1–4 providers. The purple line shows for how many subjects in each level–mode there were at least 250 FPE graduates. ‘1yr’ relates to the 2015/16 graduation cohort, ‘2yr’ relates to the 2015/16 and 2014/15 graduation cohorts and ‘3yr’ relates to the 2013/14, 2014/15 and 2015/16 graduation cohorts.

The graph shows that for full-time first degree undergraduate courses, a large majority of subjects meet our restrictions based on just a single year of data, and nearly all subjects meet them if two years of data are pooled. For other level–mode combinations, coverage is much sparser, but there are also fewer subjects for which we could hope to produce estimates, as indicated by the purple line. Across mode and level, the gain of moving from two to three pooled cohorts is typically modest compared with the gain of moving from one to two pooled cohorts. We therefore conclude that the decision to pool across two cohorts provides an appropriate balance.

6. Baseline estimates

In this section, we present estimates from our baseline model. As set out in Table 1, we produce estimates for 2,172 courses (provider–subject–level–mode combinations) that meet the minimum sample size restrictions detailed in the previous section. This means that we cover only 31% of all courses offered; however, as mostly small courses are excluded, this 31% of courses covers 80% of all full-person equivalent (FPE) students in the data (row [4] of Table 1). This share varies a lot across level–mode combinations: our estimates cover 93% of full-time first degree students, but only 40% of full-time Other UG students, 84% of full-time UG with PG students, 44% of part-time first degree students and 53% of part-time Other UG students.²⁶

Row [5] of Table 1 shows the proportion of graduates we actually use in estimation among courses for which we are producing estimates. Two-thirds or more of students are included for the full-time courses, but less than half are included for part-time students. This is mainly because part-time students tend to be older at the time of graduation, meaning we do not observe a school record for them and therefore exclude them from the estimation.

The final row of Table 1 shows the share of students in each level–mode that are on courses for which we report *and* that whose data was included in the estimation of their values. This measure is more than two-thirds for full-time first degree graduates and still 60% for full-time UG with PG graduates, but less than 30% for part-time first degree graduates and both modes for Other UG.

Figure 7 presents the *overall* distribution of our baseline earnings metric of the difference from the benchmark for each course (subject–provider combination) within each level–mode combination. As a reminder, the percentage differences from the benchmark are the adjusted β coefficients from equation (2) – see Section 5.iii for more details.

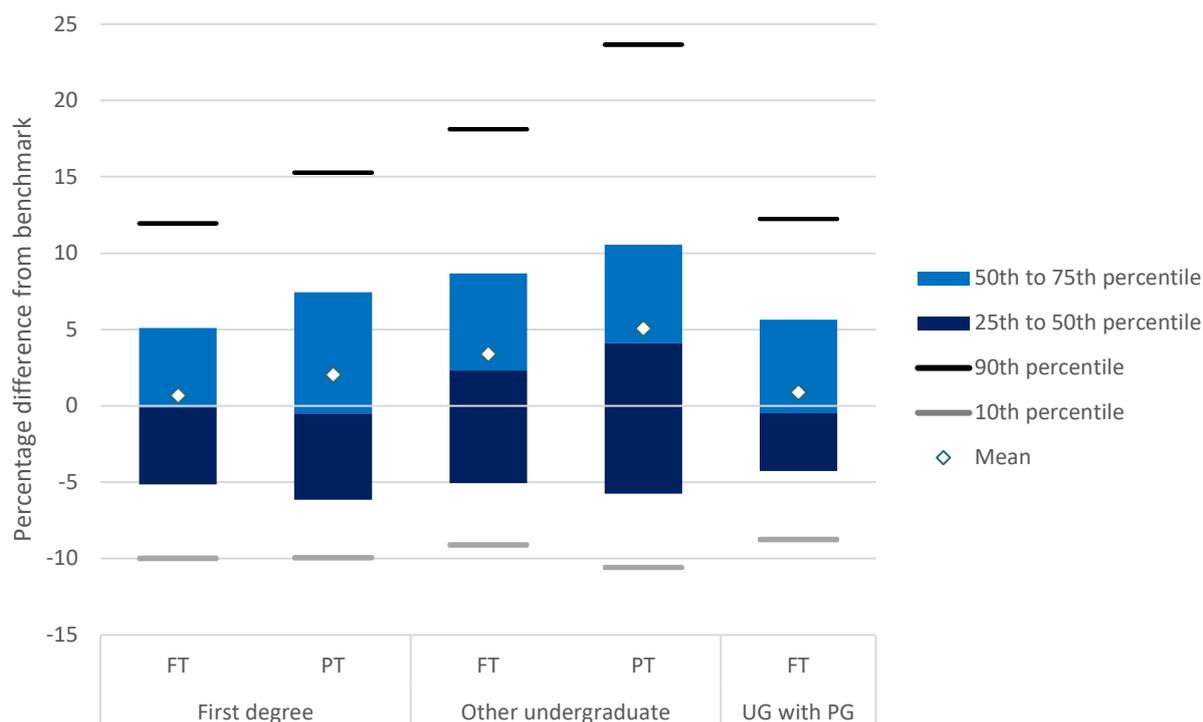
²⁶ See Table 13 in the appendix for coverage if an alternative sample size restriction of 23 FPE graduates was adopted, instead of 50. This would increase the number of FPE students on courses we report estimates for by 69,000 (13%), with the biggest increase in coverage for the Other UG FT level–mode (23,900, 69%).

Table 1: Coverage
(all restricted to buckets for which we report at least one estimate)

	All	First degree FT	First degree PT	Other UG FT	Other UG PT	UG with PG FT
[1] Total no. of courses	7,104	2,739	660	2,475	935	295
[2] No. of courses for which we report estimates (% of [1])	2,172 (30.6%)	1,632 (59.6%)	62 (9.4%)	245 (9.9%)	110 (11.8%)	123 (41.7%)
[3] Total no. of FPE students	677,500	482,900	32,900	85,800	50,900	25,000
[4] No. of FPE students on courses for which we report estimates (% of [3])	544,000 (80.3%)	447,000 (92.6%)	14,500 (44.2%)	34,600 (40.4%)	27,000 (53.1%)	20,900 (83.5%)
[5] No. of FPE students on courses for which we report estimates who are also in the sample (% of [4])	385,200 (70.8%)	331,600 (74.2%)	5,000 (34.7%)	22,400 (64.7%)	11,300 (41.7%)	14,900 (71.3%)
[6] FPE students on courses for which we report estimates who are also in the sample, as a share of total FPE students	56.9%	68.7%	15.3%	26.1%	22.1%	59.5%

Note: FPE stands for full-person equivalent (see 'Definitions' in Section 1).

Figure 7: Distribution of earnings metric by course for all level–mode combinations



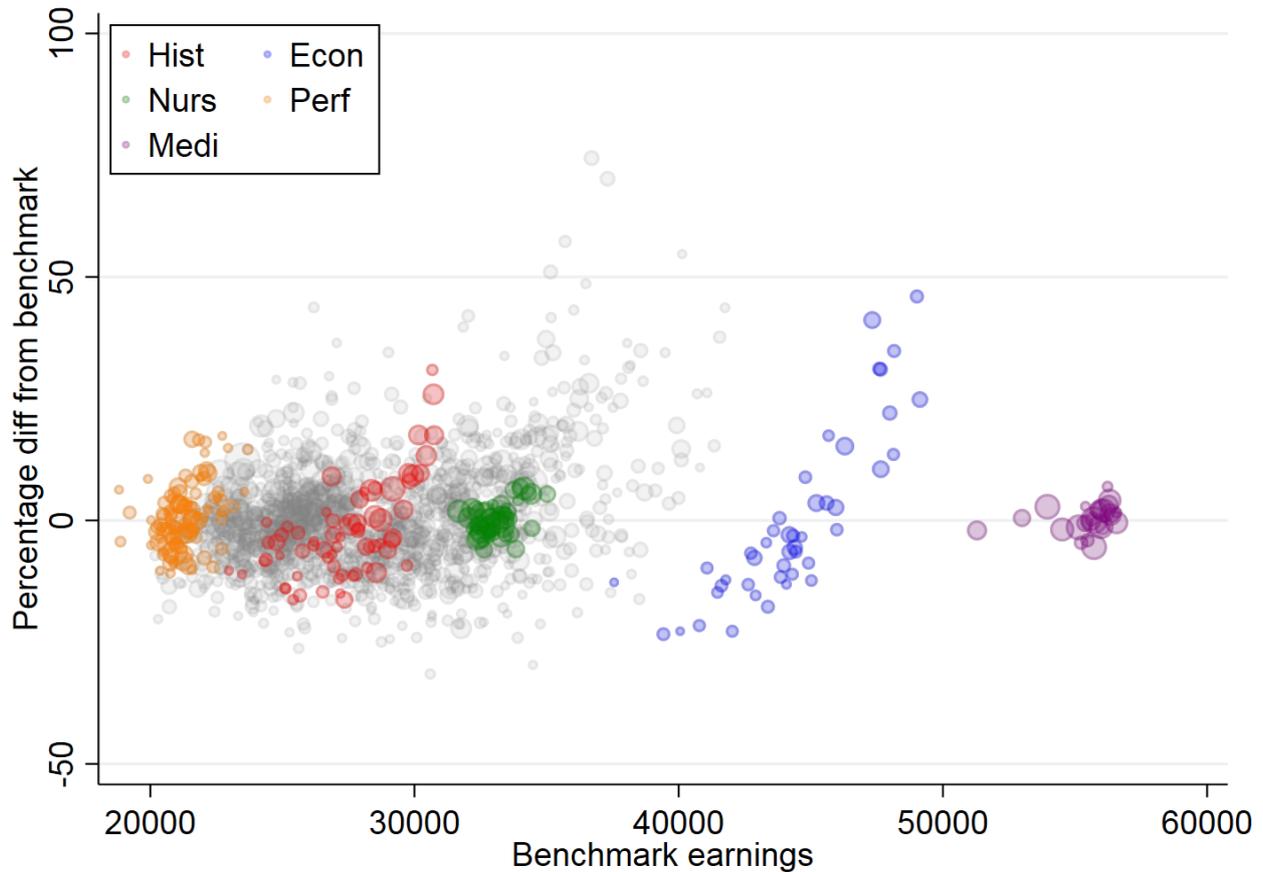
Note: Only courses for which we report estimates are included in the calculation of statistics.

For full-time first degree undergraduate courses, around half of our HEP effects are within 5 percentage points of their benchmark, but over one-fifth of courses are more than 10 percentage points above or below the benchmark. Owing largely to smaller sample sizes, the distribution of estimates is wider for other undergraduate and part-time courses. The weighted average for these level–mode combinations is substantially positive, as small courses, which on average have worse outcomes, are included in the model but not included in the calculation of the average.

Figure 8 shows some more detail for full-time first degree undergraduate courses.²⁷ Percentage differences from the benchmark are plotted against benchmark earnings estimates for all courses (as a reminder, the benchmark is the expected earnings for each institution – again, see Section 5.iii for more detail). The size of each bubble represents the number of FPE students on each course. History courses are highlighted in red, economics courses in blue, nursing courses in green, performing arts courses in orange and medicine courses in purple.

²⁷ Figure 36 in the appendix shows the equivalent for full-time other undergraduate courses, highlighting business and management, engineering, nursing and performing arts.

Figure 8: Baseline benchmark and difference from benchmark estimates, all courses, full-time first degree undergraduates



Note: Includes all courses for which we report estimates. The size of each bubble represents the number of FPE graduates from each course included in the sample. History courses are highlighted in red, economics courses in blue, nursing courses in green, performing arts courses in orange and medicine courses in purple.

From Figure 8, it becomes clear that benchmark earnings differ a lot by subject of study: compared with graduates from all subjects, for example, the benchmark earnings of performing arts graduates are low, those of economics graduates are very high and those of medicine graduates are exceptionally high. Subjects also differ in the spread of benchmark earnings, with large differences in benchmark earnings between courses for economics and quite small differences for nursing. This likely reflects the fact that nearly all nursing graduates become NHS nurses, who are paid on a common pay scale regardless of school attainment and other background characteristics.

Within each subject, percentage differences from the benchmark are centred at zero by construction. However, subjects differ in the spread of percentage differences from the benchmark: this is very high for economics and very low for nursing and medicine, suggesting that which provider a student attends matters a lot for earnings outcomes in economics but little in medicine and nursing. The latter is likely related to the fact that access to standard NHS career paths does not differ depending on the provider at which a graduate studied. For some subjects, including economics and history, benchmark

earnings and the percentage difference from benchmark are strongly correlated – an issue we return to below.

Overall, 71% of the variance in average actual earnings across courses is accounted for by the variation in benchmark earnings (see Table 15 in the appendix). Only 16% is accounted for by the variance of the earnings metric and 13% by the covariance between the benchmark and the earnings metric. This highlights how much of the differences in earnings across courses is accounted for by differences across subjects and the characteristics of course intake, and thus how important it is for regulatory purposes to benchmark courses' earnings outcomes appropriately.

In the remainder of this section, we dig deeper into these results for a single subject across providers and for a single provider across subjects. First, we show results for an example bucket of full-time first degree undergraduates in a particular subject, history. Second, we show estimates across a range of subjects for one example provider. Finally, we discuss how the results for each provider could be presented in the OfS's existing B3 dashboard.

These results are only meant as illustrations to allow the reader to visualise some of the results. We provide full results in the accompanying spreadsheet which includes all the estimates that our models produce, given the sample size restrictions we impose in the baseline case.

Results for a single subject across providers

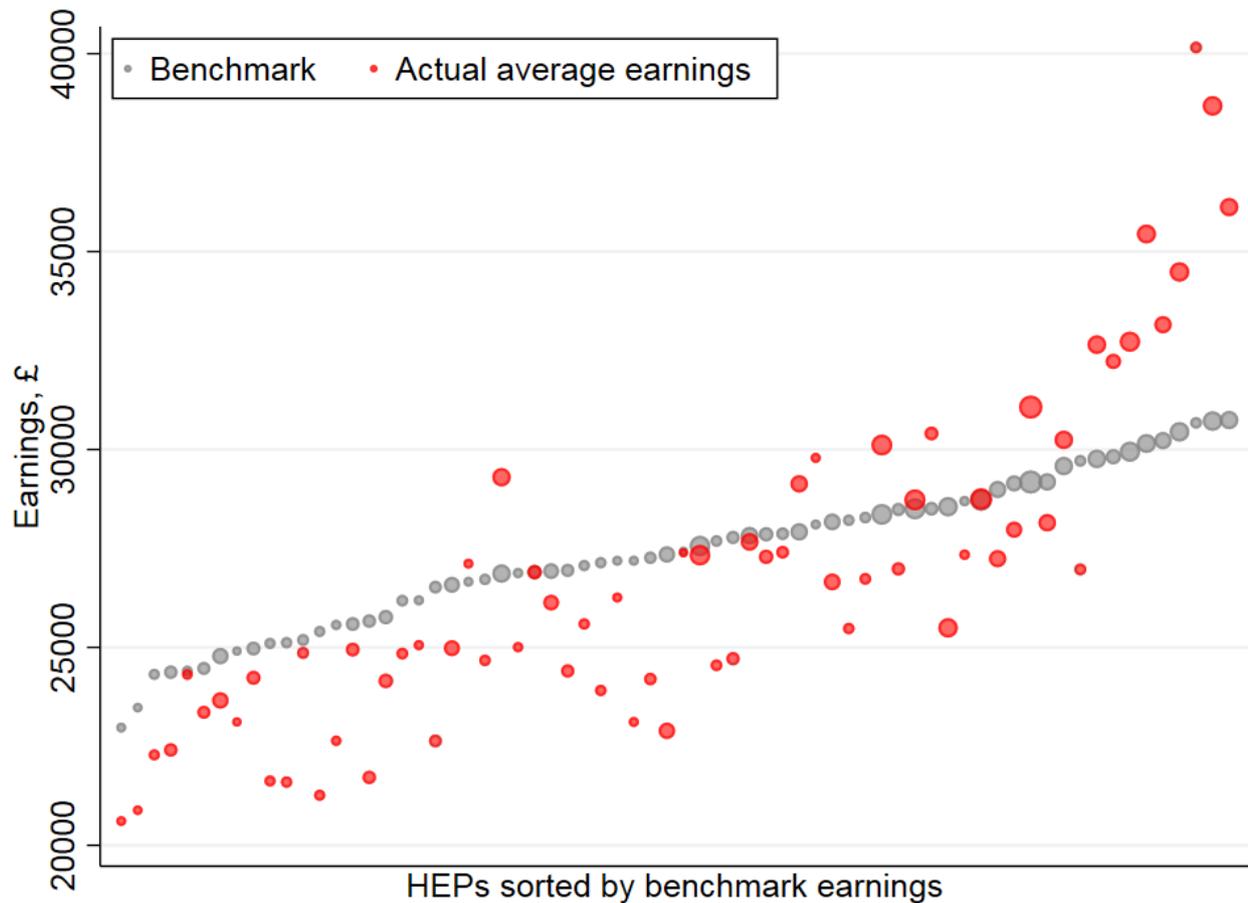
Here we present results for a single bucket – full-time first degrees in history – across all providers that meet our requirement for the minimum number of students (50 across two cohorts). This bucket is chosen because it contains many providers and a large number of students, who typically go on to have middling earnings outcomes. Like graduates from most degree subjects – and unlike graduates of a few vocational subjects such as medicine or nursing – history graduates work in a variety of roles across industries with no typical career trajectory.

Figure 9 presents benchmark estimates (in pounds) and actual average earnings by institution (both in 2021/22 prices) for history, with providers sorted in ascending order by average benchmark earnings. As discussed in Section 5.ii, these both draw upon the highest recorded earnings between the third and fifth years after graduation.

Average benchmark earnings by course vary between around £23,000 and around £30,000. The variation in actual average earnings across providers is much greater at between £20,000 and £40,000. There is a clear correlation between benchmark and actual earnings, but the relationship is far from perfect, implying that even within the same subject area, there is substantial heterogeneity in graduate earnings across providers that cannot be predicted from the students' characteristics. Providers with the

highest benchmark earnings tend to have much greater actual earnings which far exceed their benchmark earnings.

Figure 9: Average benchmark and actual earnings for full-time first degree courses in history

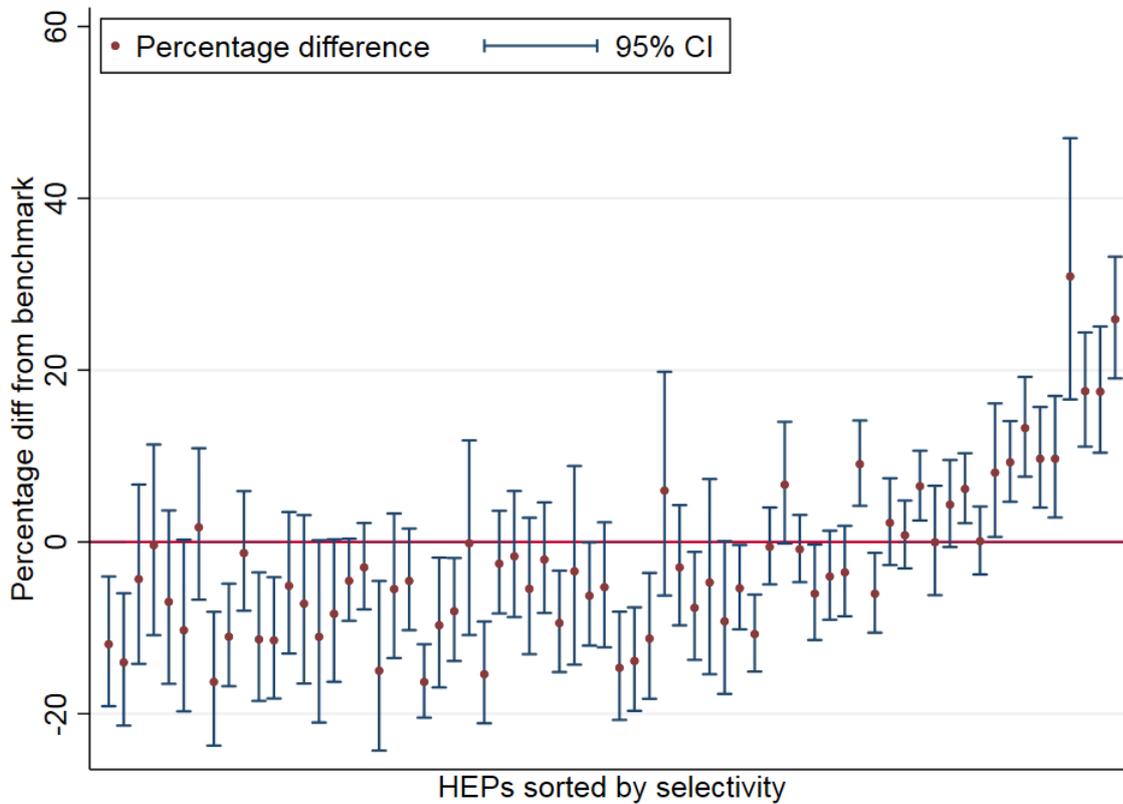
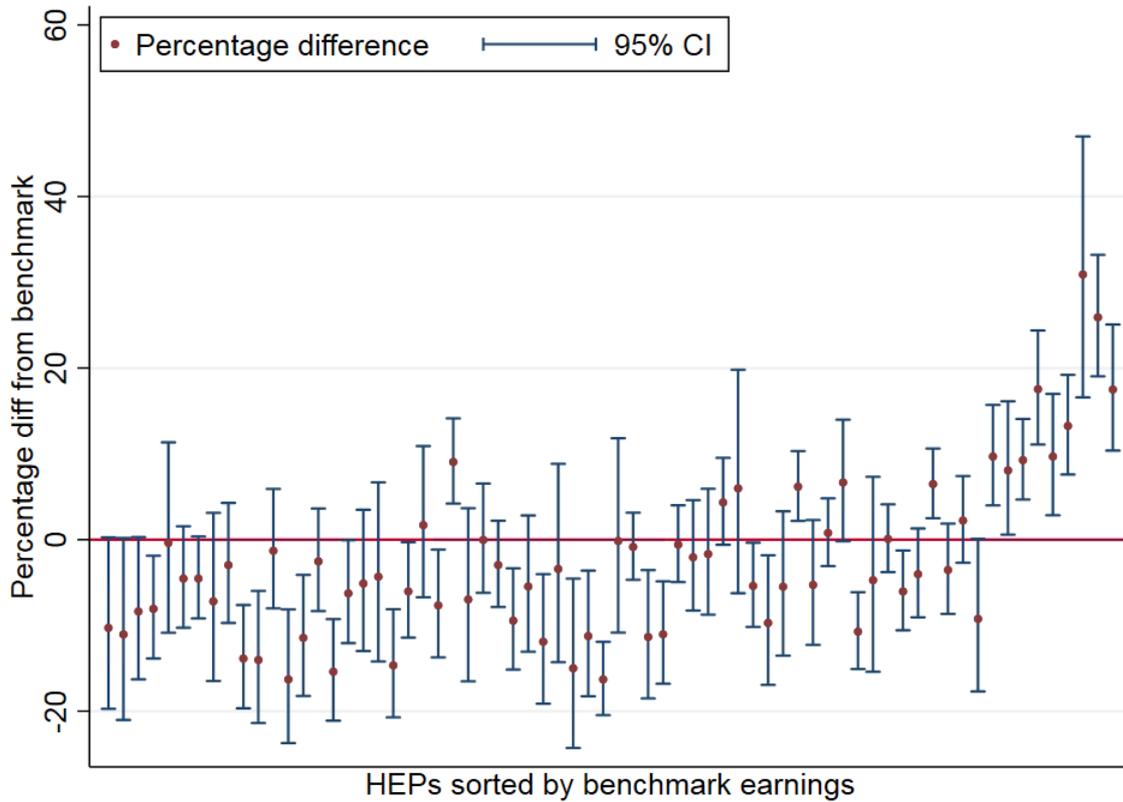


Note: The size of each bubble represents the number of FPE students on each course. 'Actual average earnings' is the exponential of mean log earnings. Figures are CPI real, in 2021/22 prices.

Figure 10 shows the percentage differences between the benchmark earnings and the actual average earnings for students on a course. These are the β_j s from equation (2) in Section 5.iii, converted from log points into percentages. Whiskers indicate 95% confidence intervals. This is our baseline earnings metric against which providers could be measured. The upper panel shows estimates sorted by benchmark earnings, and the lower panel shows estimates sorted by course selectivity, which we measure by the average GCSE score of its graduates.²⁸

²⁸ Even though universities do not usually select students based on their average GCSE scores, the average GCSE score is a common measure of university selectivity (e.g. Britton et al., 2022). The key advantage relative to Key Stage 5 scores is that GCSEs provide a common standard across students that specialise in different A level (or equivalent) subjects.

Figure 10: Percentage difference from benchmark for full-time first degree students in history



Note: Percentage differences from benchmark are the β_j s from equation (2), converted from log points into percentages. Whiskers indicate 95% confidence intervals.

Figure 10 emphasises the point that for history, this method produces large positive estimates for a small group of institutions with high benchmarks and it produces small or negative estimates for most others. Benchmark earnings are highly correlated with selectivity, suggesting that this is the result of history graduates at a few highly selective institutions enjoying high earnings even compared with graduates with similar characteristics who went to less selective institutions. As the average difference from the benchmark across graduates within a subject is zero by construction, this pushes the estimated difference below zero for many less selective providers.

This pattern is not unique to history (see the left-hand panel of Figure 27 in the appendix) or indeed to our baseline approach in this report. The same pattern of low or negative relative earnings returns outside of highly selective institutions has been found previously in academic work (Britton et al., 2022). The difference between the OfS's progression indicator and the corresponding benchmark is also correlated with selectivity (right-hand panel of Figure 27), although the correlation is somewhat weaker.

This result is not overly surprising: the most selective universities are so selective partly because they are known to offer better access to highly paid jobs even conditional on the characteristics of their intake. While it is also possible that these patterns are the result of bias arising from insufficient controls for student intake, we show in Section 8 that they are robust to many different ways of controlling for student characteristics.

However, the high relative earnings returns of highly selective institutions may well be largely due to factors that are not directly under providers' control. Selectivity – over which it could be argued the majority of providers have limited direct control – may well affect graduates' success in the labour market directly. For instance, more selective universities may have a better reputation among employers; having higher-achieving peers may lead graduates to choose higher-earning careers; and being able to draw on a network of higher-earning peers and university alumni could directly influence graduates' labour market opportunities.

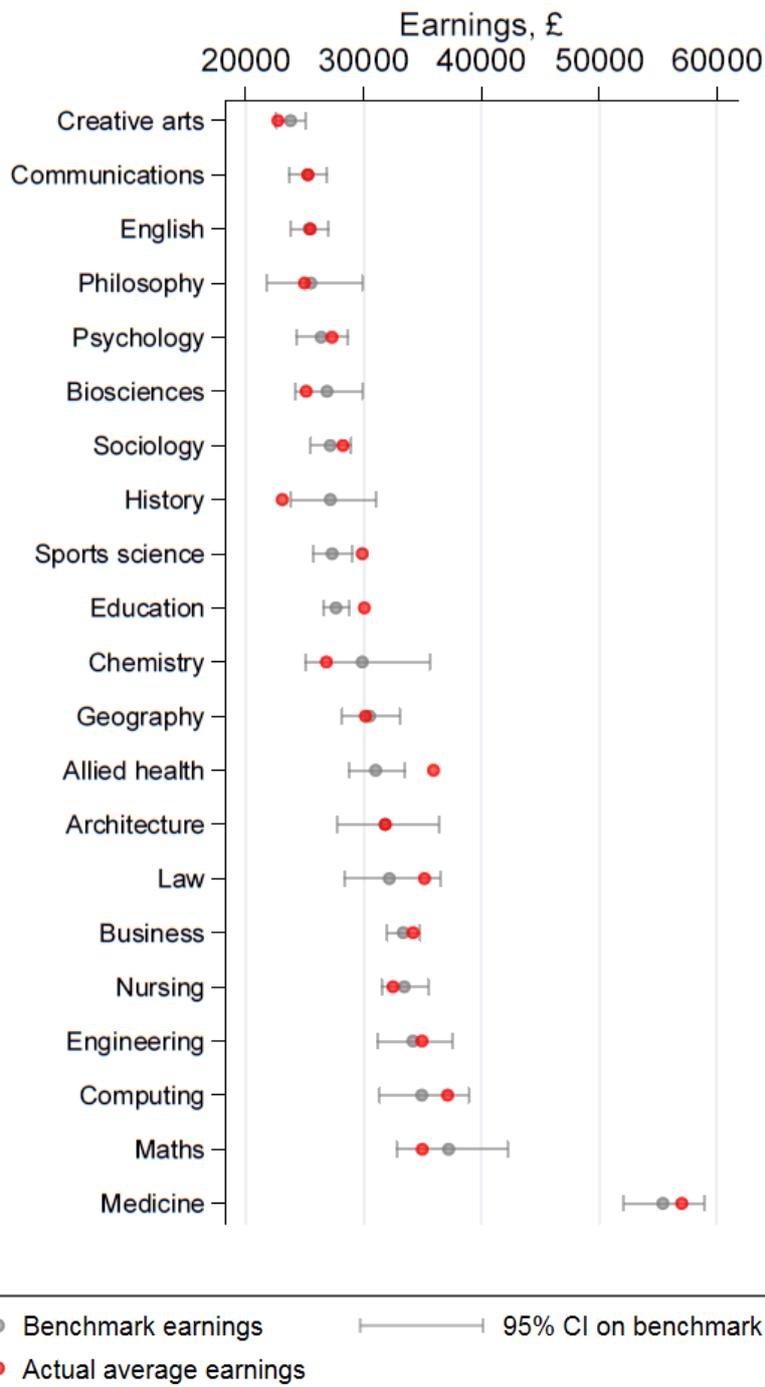
As discussed under 'Conceptual criteria' in Section 3, there are arguments for an earnings metric to abstract from such effects. However, this would require moving away from an earnings metric that targeted the full causal impact of a course on graduates' earnings; it would also be very difficult to do in practice. We set out one possible approach to this in Section 8.

Results for a single provider across subjects

This section presents analogous results for a single (unnamed) provider – a large and moderately selective university – focusing again on full-time first degrees. Figure 11 shows average benchmark and actual earnings across all subjects offered at this provider with sufficient numbers of students. Both benchmark and actual earnings are

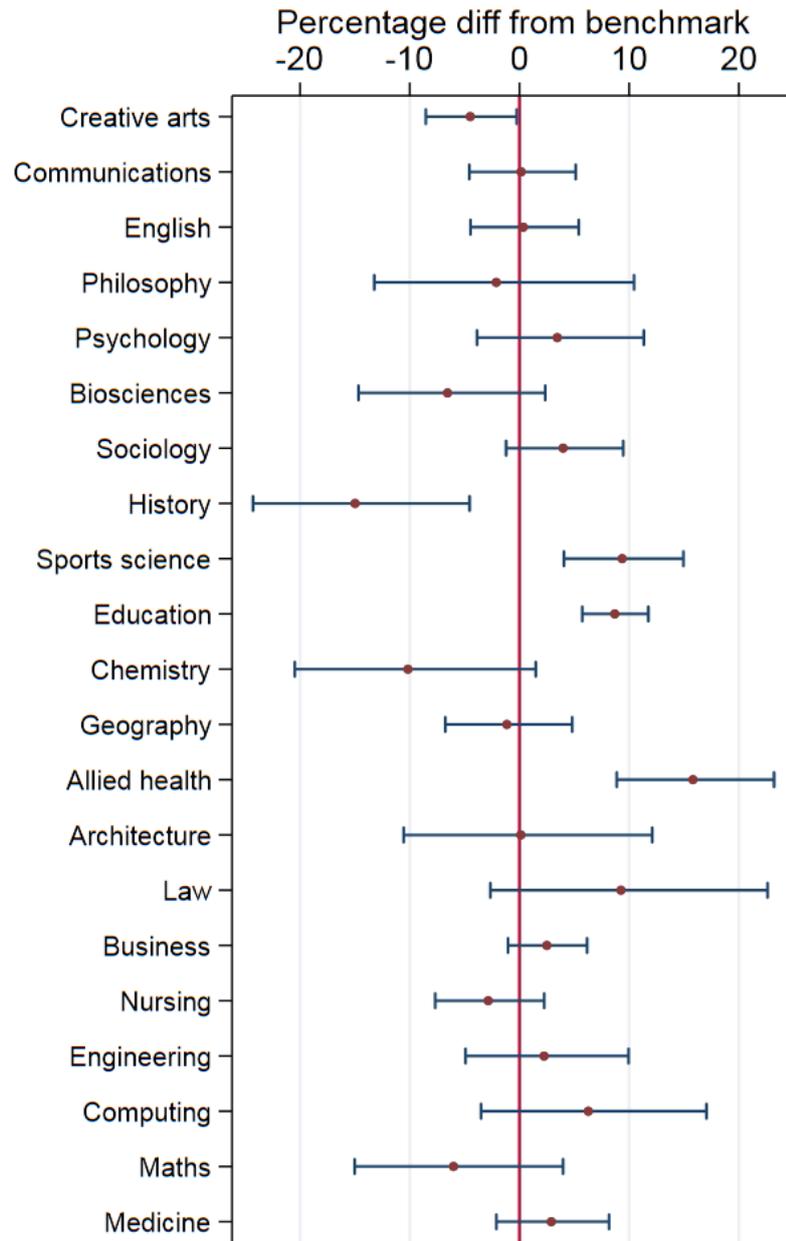
lowest for creative arts graduates at just above £20,000 and highest for medicine graduates at around £55,000 on average.

Figure 11: Average benchmark and actual earnings for full-time first degree students at an example provider



Note: 'Actual average earnings' is the exponential of mean log earnings. Whiskers indicate 95% confidence intervals on benchmark earnings.

Figure 12: Percentage difference from benchmark for full-time first degree students at an example provider



• Percentage difference — 95% CI on percentage difference

Note: Percentage differences from benchmark are the β_j s from equation (2), converted from log points into percentages. Whiskers indicate 95% confidence intervals on the differences from benchmark.

Figure 12 shows the difference from the benchmark for each of these subjects. For most subjects, the difference between actual and benchmark average earnings is not statistically significant at the 95% confidence level, meaning that in most subject areas, in terms of graduate outcomes, this university is performing about as expected. Only the estimate for history is significantly negative – suggesting history graduates from this

university are underperforming in the labour market relative to expectation – while the estimates for sports science, education and subjects allied to medicine are significantly positive, suggesting the university is outperforming expectations in these areas.

Table 2 provides the full set of results for the provider. The first two columns are the average actual and benchmark earnings as presented in Figure 11. The next three columns give the difference from the benchmark as presented in Figure 12, including the start and end points of the 95% confidence interval. Then comes the ‘full-person equivalent’ number of students included in the model for each subject (graduates from joint degrees are weighted proportionally according to the weighting assigned to each subject in the HEP records).

Table 2: Full results for an example provider

	Actual average earnings	Benchmark	Difference (ppt)		FPE graduates included	Proportion graduates included	Proportion graduates earning less than £3,000	
			Lower CI	Upper CI				
Creative arts	22,700	23,800	-4.5	-8.5	-0.2	540	75%	8%
Communications	25,300	25,300	0.2	-4.6	5.1	238	83%	7%
English	25,500	25,400	0.4	-4.5	5.4	210	85%	8%
Philosophy	25,000	25,500	-2.1	-13.2	10.5	92	79%	9%
Psychology	27,300	26,400	3.5	-3.9	11.3	67	75%	-
Biosciences	25,100	26,900	-6.6	-14.7	2.3	119	73%	11%
Sociology	28,300	27,200	4.0	-1.2	9.5	196	80%	6%
History	23,100	27,200	-15.0	-24.3	-4.5	67	75%	-
Sports science	29,900	27,300	9.4	4.1	14.9	276	84%	4%
Education	30,100	27,700	8.7	5.7	11.7	489	78%	4%
Chemistry	26,900	29,900	-10.2	-20.5	1.5	69	77%	-
Geography	30,200	30,500	-1.1	-6.8	4.8	181	83%	7%
Allied health	35,900	31,000	15.8	8.9	23.2	101	58%	-
Architecture	31,900	31,800	0.1	-10.6	12.1	113	77%	-
Law	35,200	32,200	9.3	-2.7	22.6	53	84%	-
Business	34,200	33,400	2.5	-1.0	6.2	702	79%	7%
Nursing	32,500	33,500	-2.8	-7.7	2.3	223	45%	3%
Engineering	35,000	34,200	2.2	-4.9	9.9	116	71%	8%
Computing	37,100	34,900	6.3	-3.5	17.1	119	74%	7%
Maths	35,000	37,200	-6.0	-15.0	4.0	74	79%	-
Medicine	57,000	55,400	2.9	-2.1	8.2	75	76%	-

Note: ‘Proportion graduates included’ is the share of graduates with sufficient background controls to be included in the estimation *and* at least one earnings observation of at least £3,000 three–five years after graduation. The proportion earning below £3,000 is as a share of all individuals on the course.

The penultimate column gives the share of all graduates included in the estimation, while the final column gives the share of graduates whose highest earnings do not reach £3,000 three–five years after graduation. The latter is an important driver of graduates being excluded from the estimation of course effects. However, as can be inferred from the figures in the penultimate column, a big driver of graduates being excluded from the estimation is them having insufficient background characteristics to be included. This is

usually because they are too old to match to the school records, which are only available for those who took their GCSEs in the summer of 2002 or after (see Section 4 for more on this). With the exception of nursing (which is disproportionately studied by mature students), the shares of graduates included in estimation are comfortably above response rates in the Graduate Outcomes survey. And because the shares of students who took their GCSEs before 2002 will decline, the proportion of students included in the estimation will increase over time.

Table 3 presents additional contextual information for the same provider that helps add context to the results in Table 2. The first two columns add context regarding students' working location as adults. Across subjects, only a small fraction of students are staying in the same area as this provider after graduation. A large minority in all subjects move to London, except in health-related fields and education. This is relevant information for identifying whether a course is serving specific skills needs in the local area.

Table 3: Additional contextual information for an example provider

	Proportion living in same TTWA as HEP	Proportion living in London TTWA	Proportion with any SA earnings	Proportion of all earnings from SA	Average GCSE point score (standardised)	Selectivity percentile rank within bucket
Creative arts	13.7%	38.4%	23.8%	21.6%	0.67	0.89
Communications	10.1%	33.1%	8.6%	7.8%	0.36	0.31
English	17.1%	24.1%	5.6%	4.8%	0.62	0.48
Philosophy	17.5%	31.1%	8.3%	7.5%	0.48	0.02
Psychology	14.6%	30.0%	5.3%	5.1%	0.57	0.41
Biosciences	8.5%	31.8%	3.9%	3.5%	0.83	0.57
Sociology	17.8%	30.1%	2.8%	2.6%	0.49	0.58
History	20.3%	44.5%	6.3%	5.5%	0.54	0.27
Sports science	3.5%	31.2%	7.8%	6.9%	0.54	0.77
Education	13.5%	18.0%	2.8%	2.2%	0.64	0.91
Chemistry	7.2%	39.0%	2.3%	1.7%	0.83	0.38
Geography	13.1%	28.8%	1.1%	1.1%	0.62	0.35
Allied health	9.9%	17.8%	13.8%	12.2%	0.77	0.57
Architecture	14.3%	41.1%	10.6%	9.6%	0.56	0.52
Law	15.4%	35.9%	2.6%	2.7%	0.63	0.54
Business	9.4%	42.3%	5.3%	4.7%	0.53	0.61
Nursing	16.8%	12.7%	1.4%	1.1%	0.55	0.72
Engineering	12.3%	25.0%	7.0%	6.3%	0.57	0.48
Computing	10.6%	34.0%	8.4%	7.3%	0.44	0.56
Maths	12.2%	43.5%	0.0%	0.0%	0.76	0.21
Medicine	14.5%	18.8%	4.1%	2.8%	1.32	0.56

Note: Whether a graduate is living in a given travel-to-work area (TTWA), and whether they have any earnings from self-assessment (SA), are based on their highest-earning tax year between three and five years after graduation. 'London TTWA' includes both 'London' and 'Slough and Heathrow'.

The third and fourth columns provide context on whether graduates from these courses are employed or self-employed after graduation. At this provider, only creative arts graduates had substantial earnings (although still less than a quarter of total earnings) that they declared as self-employment earnings or partnership profits on their self-

assessment (SA) tax returns. This is relevant contextual information, as those setting up their own businesses may have comparatively lower earnings for a period.

The final two columns provide context on course selectivity. The penultimate column gives the average standardised GCSE score of the course's graduates; for example, a score of 1 would mean that graduates of this course scored on average one standard deviation above the overall mean. The final column gives the selectivity rank within bucket (subject–level–mode category). This provides an indication how selective the course is relative to other course in the same subject area.

Relative to other courses in the same subject area, courses at this provider range from the highly selective (education) to the least selective (philosophy). In absolute terms, most courses are moderately selective except for medicine, which is very selective. It should be noted, however, that selectivity here is defined by the average GCSE score, while courses will in practice select on other student characteristics.

The contextual information presented in Table 3 is merely meant to be indicative of the type of contextual information that could be useful. Other indicators might have been chosen, and indeed other types of information could be helpful. For instance, the share of graduates entering a particular industry might be useful context for courses serving that industry.

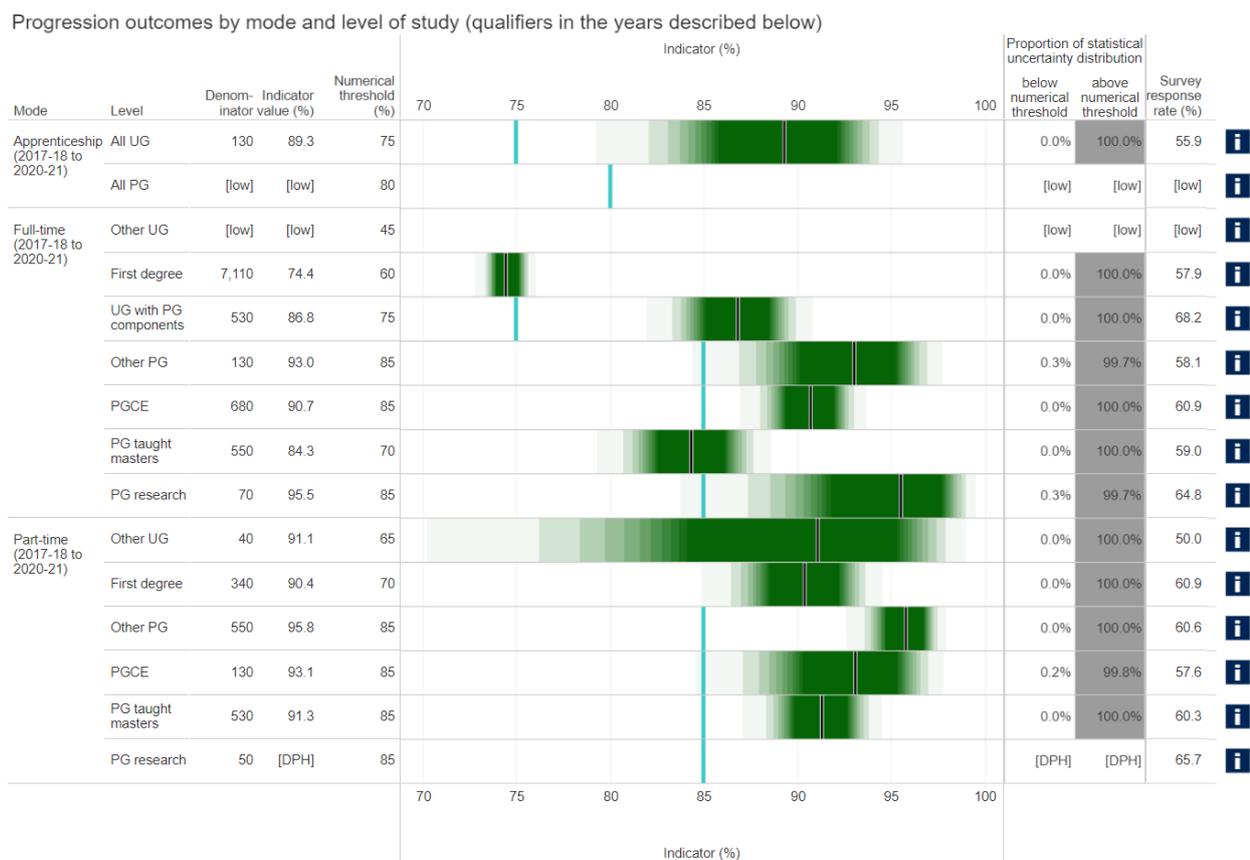
Integrating our results with the OfS dashboards

This section discusses how these results could be integrated into the OfS's existing dashboards, focusing specifically on the 'B3' metrics. This dashboard on the OfS website²⁹ visualises the results for the B3 metrics for all OfS-registered providers, both at the provider level within each mode and level of study and broken down as 'split indicators'. Here we show the results for the progression metric, as this is the B3 metric that is closest to an earnings metric, but the continuation and completion metrics are displayed in the same format.

Figure 13 shows the overview page for the progression metric from the OfS B3 dashboard for the same example provider whose results on our baseline metric were presented above. For each combination of level and mode – except where results are suppressed due to a small number of observations – an *overall* average indicator value is shown together with a fixed numerical threshold. Green error bands visualise the statistical uncertainty around each indicator value.

²⁹ <https://www.officeforstudents.org.uk/data-and-analysis/student-outcomes-data-dashboard/data-dashboard/>.

Figure 13: Overview of progression outcomes for an example provider, OfS B3 dashboard



Source: www.officeforstudents.org.uk.

The first thing to note about integrating our estimates into this framework is that we only cover five combinations of mode and level: full-time and part-time first degree, full-time and part-time Other UG, and full-time UG with PG components. This is largely due to the limited scope of this report. Future work could extend similar methods to other mode and level combinations.

More fundamentally, the actual average indicator values on the OfS B3 dashboard do not adjust for providers' intake. As we have argued in Section 3, we do not recommend reporting earnings in this way, as average graduate earnings will in many cases be substantially affected by factors that are unrelated to their higher education courses. At the provider level, these actual earnings would additionally be strongly affected by the mix of subjects taught at a provider (for instance, the provider-level results for the example provider discussed in the previous subsection would appear substantially better on account of having a medical school).

If provider-level results were to be reported, we would recommend for them to be reported as a percentage difference from the benchmark for each provider, with or without a numerical threshold. While our results focus on the course (subject-provider) level within each level-mode combination, it is straightforward to aggregate them up to

the provider level. For each provider, the difference from the overall benchmark $\hat{\beta}_j^*$ is given by the weighted average of subject-specific differences, so

$$\hat{\beta}_j^* = \frac{\sum_s N_{js} \hat{\beta}_{js}}{\sum_s N_{js}}$$

where N_{js} is the full-person equivalent number of students studying subject s in level-mode j . As the course-level estimates are based on independent samples, the standard error of the provider-level estimate can be approximated by³⁰

$$SE(\hat{\beta}_j^*) = \frac{\sqrt{\sum_s (N_{js} SE(\hat{\beta}_{js}))^2}}{\sum_s N_{js}}$$

As earnings are approximately log-normally distributed and a Central Limit Theorem applies, the distribution of $\hat{\beta}_j^*$ will be very close to Normal, so similar error bands would be straightforward to produce.

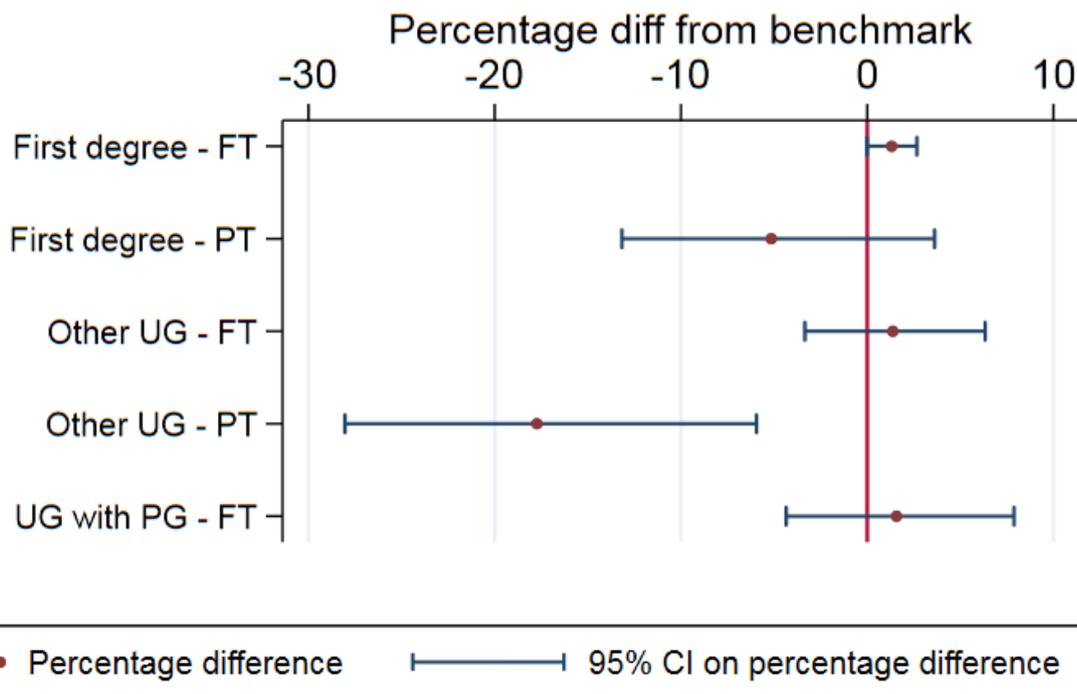
Figure 14 shows how the results for our example provider could be reported, showing overall estimates for each level-mode, with 95% confidence intervals on the estimates.³¹ The confidence intervals are much narrower on full-time first degrees because there are much larger sample sizes. Consequently, the relatively small estimate of 1.3% above the benchmark suggests overperformance of the university on full-time first degrees relative to expectation, but this is (just) below the threshold for significance. The estimates for part-time first degree, full-time Other UG and UG with PG are also statistically not significantly different from zero. The estimate for part-time Other UG is based on a relatively small number of students, but nevertheless is significantly below the benchmark, suggesting underperformance of the university on these degrees relative to expectations. To clarify, even though these estimates are presented at the provider level, they still reflect outcomes relative to the average across providers within each subject.

Figure 13 also shows supplementary information provided in the table columns in addition to the indicator visualisation. All have direct analogues in our baseline approach. ‘Denominator’ is the equivalent of ‘FPE graduates included’ in Table 2. The ‘proportions of statistical uncertainty distribution’ below and above a numerical threshold would be straightforward to calculate given the standard error and the properties of the Normal distribution. The ‘survey response rate’ corresponds to the proportion of graduates included from Table 2.

³⁰ This will slightly understate the standard error of the provider-level estimate as a result of covariance between course-level estimates induced by students taking joint honours and thus being included in both regressions (at full-person equivalent weight).

³¹ Provider-level estimates for full-time first degree courses for all providers are presented in Figure 25 in the appendix.

Figure 14: Provider-level differences from benchmark at an example provider



Note: Whiskers indicate 95% confidence intervals on percentage difference.

Figure 15: Detailed progression outcomes for an example provider, OfS B3 dashboard, indicator values

Progression outcomes (2017-18 to 2020-21 qualifiers) by subject, student and study characteristic [▼ Show proportion of negative outcomes with interim study](#)

Split indicator type	Split indicator	Denominator	Indicator value (%)	Indicator (%)						Proportion of statistical uncertainty distribution		Benchmark value (%)	Contribution to own benchmark (%)	Survey response rate (%)	i		
				40	50	60	70	80	90	100	below numerical threshold					above numerical threshold	
Overall indicator	Overall	7,110	74.4									0.0%	100.0%	73.0	5.7	57.9	i
Time series	Year 1 (2017-18)	1,820	77.3									0.0%	100.0%	74.5	5.6	58.8	i
	Year 2 (2018-19)	1,950	72.6									0.0%	100.0%	71.2	5.4	58.6	i
	Year 3 (2019-20)	1,820	72.5									0.0%	100.0%	71.9	6.0	57.7	i
	Year 4 (2020-21)	1,520	75.6									0.0%	100.0%	74.9	5.8	56.2	i
Type of partnership	Taught by this provider	7,040	74.7									0.0%	100.0%	73.1	5.7	58.1	i
	Only registered by this provider (subcontracted out)	70	51.4									92.7%	7.3%	59.8	58.7	43.7	i
Course type	First degree with integrated foundation year	130	74.4									0.0%	100.0%	73.7	19.1	68.4	i
Subject: Business and management	Business and management	970	71.1									0.0%	100.0%	67.6	2.6	55.2	i
Subject: Design, and creative and performing arts	Creative arts and design	950	67.5									0.0%	100.0%	69.6	4.5	54.1	i
	Performing arts	20	78.2									3.1%	96.9%	68.7	3.0	45.3	i
Subject: Education and teaching	Education and teaching	600	84.4									0.0%	100.0%	75.4	6.0	61.7	i
Subject: Engineering, technology and computing	Computing	250	83.3									0.0%	100.0%	79.7	3.1	61.1	i
	Engineering	270	75.8									0.0%	100.0%	74.3	5.7	64.5	i

Source: www.officeforstudents.org.uk.

Figure 15 shows the detailed view on the OfS dashboard for the progression outcomes at the same example provider for full-time first degree students only (truncated at the bottom). The overall indicator is reproduced at the top, with the remaining rows showing various ‘split indicators’, including by subject. Again, actual indicator values are shown together with a fixed numerical threshold. The numerical threshold is the same for all split indicators, as it is only differentiated by level–mode groups.

Again, these are actual average indicator values that do not adjust for the provider’s intake. We would not recommend displaying information on actual average earnings in this way on a revised B3 dashboard. However, we do think that showing average actual earnings in a separate column would provide helpful context.

Two new columns in this view of the OfS dashboard are the ‘benchmark value’ and the ‘contribution to own benchmark’. While there is no equivalent to the latter under our baseline approach, our earnings benchmark could be calculated by split indicator as well. This would simply be the within-group average of the predicted values for the subject-specific models for all students (weighted by their full-person equivalents). This average could again be converted into pounds by applying the exponential function.

Figure 16: Detailed progression outcomes for an example provider, OfS B3 dashboard, differences from benchmark

Progression outcomes (2017-18 to 2020-21 qualifiers) by subject, student and study characteristic

Split indicator type	Split indicator	Denominator	Indicator value (%)	Difference from benchmark (ppt)						Proportion of statistical uncertainty distribution		Benchmark value (%)	Contribution to own benchmark (%)	Survey response rate (%)		
				-30	-20	-10	0	10	20	30	below benchmark				above benchmark	
Overall indicator	Overall	7,110	74.4									0.2%	99.8%	73.0	5.7	57.9
Time series	Year 1 (2017-18)	1,820	77.3									0.1%	99.9%	74.5	5.6	58.8
	Year 2 (2018-19)	1,950	72.6									6.7%	93.3%	71.2	5.4	58.6
	Year 3 (2019-20)	1,820	72.5									25.3%	74.7%	71.9	6.0	57.7
	Year 4 (2020-21)	1,520	75.6									25.2%	74.8%	74.9	5.8	56.2
Type of partnership	Taught by this provider	7,040	74.7									0.1%	99.9%	73.1	5.7	58.1
	Only registered by this provider (subcontracted out)	70	51.4									99.2%	0.8%	59.8	58.7	43.7
Course type	First degree with integrated foundation year	130	74.4									40.9%	59.1%	73.7	19.1	68.4
Subject: Business and management	Business and management	970	71.1									0.8%	99.2%	67.6	2.6	55.2
Subject: Design, and creative and performing arts	Creative arts and design	950	67.5									92.3%	7.7%	69.6	4.5	54.1
	Performing arts	20	78.2									12.2%	87.8%	68.7	3.0	45.3
Subject: Education and teaching	Education and teaching	600	84.4									0.0%	100.0%	75.4	6.0	61.7
Subject: Engineering, technology and computing	Computing	250	83.3									6.0%	94.0%	79.7	3.1	61.1
	Engineering	270	75.8									25.5%	74.5%	74.3	5.7	64.5

Source: www.officeforstudents.org.uk.

Finally, Figure 16 shows a third view of the progression outcomes for an example provider available on the B3 dashboard, which presents the split indicators relative to the OfS's existing benchmark value. We would recommend for this to be the default view for an earnings metric. As detailed above, our baseline approach provides natural equivalents for all supplementary columns in this view apart from the 'contribution to own benchmark', which we would recommend leaving out.

7. Assessing the baseline metric

This section assesses the performance of the baseline earnings metric. We first discuss what we can learn about its performance against our criteria for a regulatory earnings metric without comparing against alternatives. Second, we compare the estimates from our baseline earnings metric with the existing OfS progression metric. Section 8 discusses the performance of variants of and alternatives to our baseline earnings metric.

Performance against our criteria for a regulatory earnings metric

The two conceptual criteria developed in Section 3 are, first, that the metric should reflect the *impact* a course has on the earnings of its graduates and, second, that it should only capture the influence of factors that providers can *control*. As discussed in that section, these criteria are largely compatible, but there remains a trade-off between them. The target for the baseline earnings metric – the causal effect of a course on its students, with the counterfactual being the average course in the same level–mode in the same subject area – strikes a balance between these two concerns. It aims at the impact of a course on graduate earnings, albeit not relative to the most intuitive counterfactual; and it aims to abstract from factors that providers cannot control, insofar as these are at the level of the individual student.

However, as was apparent in our baseline estimates for history shown in Section 6, earnings outcomes of students may substantially depend on factors at the provider level that providers themselves cannot directly control. The most important of these is provider selectivity, which arguably is something that providers largely need to take as given.

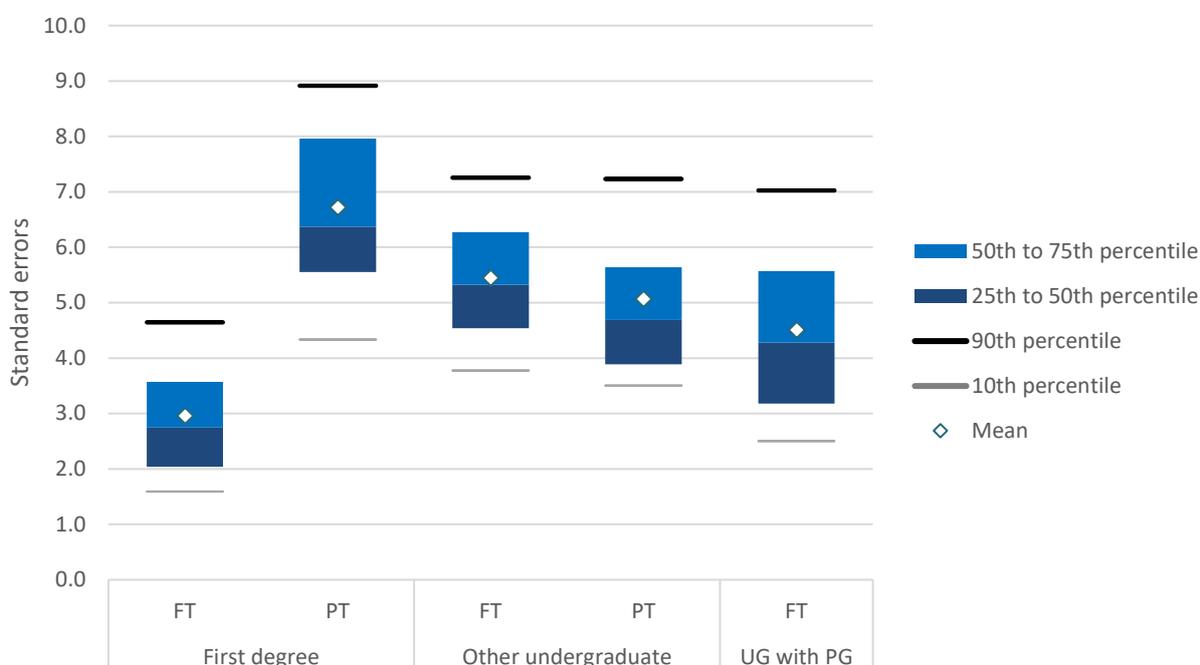
Turning to the practical criteria outlined in Section 3, this concern is also closely connected to **fairness across providers**. As argued in that section, the main question on that score is whether the earnings metric captures the intended causal effect of courses on earnings with minimal bias. While the true bias is unknowable, it is notable that our models account for a substantial share of the differences in actual earnings outcomes between courses. Reflecting these differences, benchmark earnings estimates range from below £20,000 to more than £50,000 per year. As shown in Section 8, the estimates are also robust to various changes in specification.

However, even if the baseline earnings metric recovered the true causal effect, the resulting estimates may still not be seen as fair. This is because providers at the top end of the selectivity spectrum still have large positive estimates on average, suggesting that adjusting for intake characteristics is not enough to fully capture the advantages graduates of these providers enjoy in the labour market. By construction, this means that other providers will have negative estimates on average, even though they arguably have little chance of replicating the success of the most selective providers. As argued in

Section 3, addressing this concern would require changing the target of the earnings metric, which we explore in Section 8.³²

Our second practical criterion is **accuracy**. As shown in Figure 17, standard errors for slightly more than half of all full-time first degree courses (weighted by the number of students) are below 3 percentage points, implying that for these courses, deviations from the benchmark of more than 6 percentage points are very unlikely to arise by chance (assuming that the statistical model reflects the true data-generating process). More than 90% of courses (weighted by the number of students) have standard errors below 5 percentage points. Overall, this level of stated statistical uncertainty is slightly larger than but still broadly similar to that for the existing OfS progression metric (see ‘Comparison with the existing OfS progression metric’ below).

Figure 17: Distribution of standard errors by level–mode, course-level estimates



Note: Only courses for which we report estimates are included in the calculation of statistics.

These numbers are encouraging, as they imply that economically significant differences from the benchmark will mostly also be statistically significant. The situation is very different for other level–mode combinations, where standard errors are much larger. Given the scale of statistical uncertainty, our proposed earnings metric will contain much less useful information for those courses.

A less formal, but likely more robust, measure of the accuracy of our earnings metric is how much it changes across cohorts. Specifically, we examine the weighted average

³² Similarly, one might be concerned about an unfair advantage for providers located in or near strong labour markets. Empirically, provider selectivity appears to play a larger role than provider location, which is why we emphasise it here.

absolute value of the change in the earnings metric across non-overlapping two-cohort intervals in our sample. Comparing the estimates for the 2014/15 and 2015/16 cohorts with the estimates for the 2012/13 and 2013/14 cohorts, the average absolute change is 3.5 percentage points for full-time first degree courses; comparing with the two cohorts before that, it is 3.7 percentage points. These small differences across cohorts suggest that our earnings metric is relatively stable over time. The magnitude of these differences, being only slightly higher than the average standard error for full-time first degree courses, indicates that the observed changes are likely due to a combination of statistical variability and gradual shifts in course performance, rather than significant fluctuations or measurement errors.³³

Based on these results, we are satisfied that for full-time first degree undergraduate courses, the baseline earnings metric is sufficiently accurate for use in regulation. However, we are much less confident for other level–mode combinations. Estimates for these level–mode combinations are substantially less robust than for full-time first degrees, partly due to the much lower number of students taught, and partly due to the larger share of mature students for whom school records are unavailable (see Table 1 in Section 6).

We would therefore advise caution in reporting any earnings metrics for part-time and Other UG courses. In our view, there is a good case not to proceed with earnings metrics for these level–mode combinations for the time being (full-time UG with PG degrees could reasonably be integrated into the full-time first degree category). In Section 8, we therefore focus on the sensitivity estimates for full-time first degree courses (full results are provided in the accompanying spreadsheet).

It is, however, worth noting that this focus on ‘standard’ undergraduate degrees would come at a cost. While the share of students on these courses is relatively small, they make up more than half of all courses (and 19% of courses for which we report estimates). A regulator may in fact be especially concerned about the outcomes of smaller providers offering less common types of courses such as part-time or Other UG degrees. There is an inescapable trade-off here between reporting a less robust estimate or no estimate at all, on which the OfS would need to take a stance.³⁴

The case for including part-time and Other UG courses may well improve over time if the number of students whose data can be used increases. Other UG courses may well become a larger part of higher education provision in the future – for instance, if Higher

³³ Unfortunately, the same exercise is not meaningful for other level–mode combinations: given our sample restrictions and changes in provision, of those courses in other categories for which we would report estimates for the 2014/15 and 2015/16 cohorts, we only report estimates for 55% of courses for the cohorts two years and four years earlier.

³⁴ This trade-off also applies to the existing B3 metrics but is perhaps more acute in our case due to our emphasis on course-specific benchmarks and using information from school data (which is unavailable for older students).

Technical Qualifications funded by the Lifelong Learning Entitlement become popular.³⁵ Furthermore, coverage will improve over time as school records will be available for ever older mature students.³⁶

Our third practical criterion is **positive incentives for providers**. There are two concerns here: that an earnings metric could distort incentives on student selection and that it could distort the incentives for the types of course offered. Controlling for nearly all of the characteristics of students that providers can observe means that the first concern is unlikely to be a major issue for our baseline metric. As all estimates are relative to courses in the same subject, our baseline earnings metric also largely addresses the second concern.

However, there are two ways in which incentives for providers to offer different courses may still be distorted. One is through the influence of factors that providers cannot control – most importantly, selectivity – on earnings outcomes. For example, if provider selectivity has a large independent effect on earnings in some subjects but not in others, this could bias less selective providers towards offering subjects such as nursing, where provider selectivity is less important for earnings, and against subjects such as economics, where it is more important. As detailed in Section 8, we suggest guarding against this by calculating ‘selectivity-adjusted benchmarks’ to be displayed in a separate column alongside the main estimates.

The other possible distortion is that, to the extent that there are differences in returns between detailed subjects within each CAH2 subject, institutions will be incentivised to shift provision towards the higher-return ones. For instance, within creative arts, graphic design might have a higher labour market return than fine art. One possible way of addressing this issue, which we explore in Section 8.vi, is to add controls for CAH3 category (167 subject groups) to the model. Whether or not this variation is adopted, we believe that this potential distortion will be substantially ameliorated by the OfS’s approach to regulation, which requires taking into account the particular shape of provision at each provider.

The fourth practical criterion is **regulatory practicality**. As shown in Section 6, our proposed earnings metric fits in well with the current B3 dashboard, and we have adopted the same definitions as the OfS as far as was practicable, including for defining subject groups (like the OfS, we use the CAH2 classification and we aggregate Celtic studies with the languages and area studies grouping). As we have argued in Section 6,

³⁵ A different aspect of the Lifelong Learning Entitlement, modular provision, will be a challenge for an earnings metric and for any other student outcome metric as it will no longer be obvious to which provider a given graduate’s outcomes should be attributed. We do not discuss this issue here, as it raises no specific issues with regard to an earnings metric.

³⁶ For example, by the end of the 2028/29 academic year, data for our baseline earnings metric should be available up to the 2021 and 2022 graduation cohorts. Among these cohorts, the youngest graduates with typical school careers who did their GCSEs before 2002 were 36 or older when they graduated. In other words, the baseline earnings metric would then capture all graduates up to age 35, which will be a large majority even for part-time courses.

the main obstacle to integrating our baseline estimates into the existing B3 dashboard would be that displaying actual average earnings for a provider compared with an absolute numerical threshold as the main indicator would not be advisable. At least for the earnings metric, the default view of the B3 dashboard would therefore need to be ‘difference from the benchmark’, which currently only plays a subsidiary role.

Our fifth practical criterion is **timeliness**. Our baseline earnings metric relies on earnings between three and five years after graduation and requires two cohorts, and there is some lag in data being made available. This means our baseline earnings metric will not be available for regulation until seven years after the first of our two cohorts have graduated, which will typically be ten years after those students entered university. This long lag could reasonably be shortened to six or even five years after graduation by looking at earnings sooner after graduation. But this would come with costs in terms of accuracy and the extent to which earnings are representative of lifetime earnings (potential). Policymakers face a multifaceted trade-off here; using two cohorts and earnings three–five years after graduation would be one of several reasonable choices.

Our final criterion is **transparency**. Our baseline metric scores highly on this criterion: it aligns closely with the existing economics literature on returns to higher education and, compared with the existing OfS progression benchmarks, the model is relatively simple. Full details of the variables that enter the baseline model are given in Table 12 in the appendix. One difference between the baseline earnings metric and the existing OfS model for calculating progression benchmarks is that the baseline earnings metric relies on the assumption that GCSE scores and (log) earnings have a linear relationship. However, we show in Section 8.vii that this is not a crucial assumption: a variation of the baseline model using GCSE score bins instead of continuous GCSE scores produces very similar estimates, and we take no view on which approach is preferable.

Comparison with the existing OfS progression metric

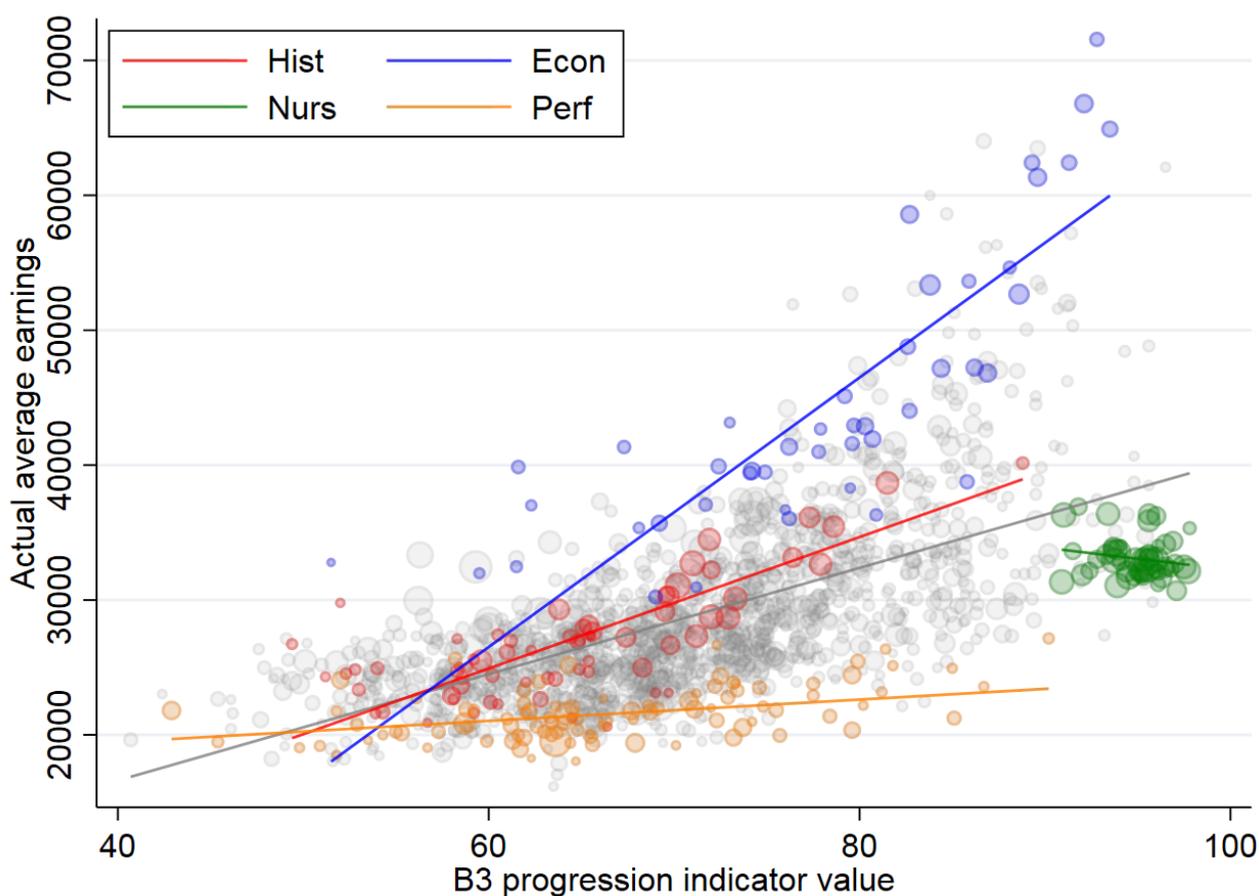
This section compares our estimates against the existing OfS progression metric. This is a useful sense-check for an earnings metric. We would expect raw average earnings to be highly correlated with the OfS progression indicator, as both are measures of labour market success. We would also expect the differences from the benchmarks to be correlated, as both are meant to be measuring providers’ success, or otherwise, in equipping their graduates with the skills necessary to succeed in the labour market.

However, we would not expect either pair of measures to be anywhere near perfectly correlated. Average graduate earnings and the average share of graduates with a ‘positive outcome’ are two quite different measures of labour market success. Even on the same measure, substantial differences could be expected, as the estimates relate to different cohorts of students.

Figure 18 plots average actual (three-year maximum) earnings and the actual OfS progression indicator value, by course. Earnings are for the 2014/15 and 2015/16 graduation cohorts and progression indicator values are for the 2018/19 to 2021/22 graduation cohorts. The size of each bubble represents the number of full-person equivalent (FPE) students on each course.³⁷ History courses are highlighted in red, economics courses in blue, nursing courses in green and performing arts courses in orange (each with a linear line of best fit included).

The overall correlation between the B3 progression indicator value and actual average earnings of students is substantial at 0.62. The relationship varies a lot by subject. For history, the B3 progression indicator and actual average earnings have a high correlation of 0.84, with the slope in line with the relationship between the two indicators across all subjects.

Figure 18: Actual average earnings and OfS progression indicator value, by course



Note: Includes all courses for which we and the OfS both report estimates. The size of each bubble represents the geometric mean of the number of FPE students on each course across the OfS data and our data. History courses are highlighted in red, economics courses in blue, nursing courses in green and performing arts courses in orange.

³⁷ More precisely, for each course-level estimate, it is the geometric mean of the FPE sample size in the OfS sample and in our sample.

Economics has a much steeper slope, indicating that earnings outcomes are much higher for economics courses that have high progression indicators. The opposite is true for performing arts, for which the slope is very shallow; in this subject, a high B3 progression indicator only corresponds to minimally higher earnings. These different slopes create sharply different average earnings by subject among courses with high B3 progression indicators. Finally, nursing is a special case: all nursing courses feature similar average earnings of £30,000–£37,000 and all feature exceptionally strong progression outcomes. Within nursing courses, there is essentially no relationship between average earnings and the progression indicator.

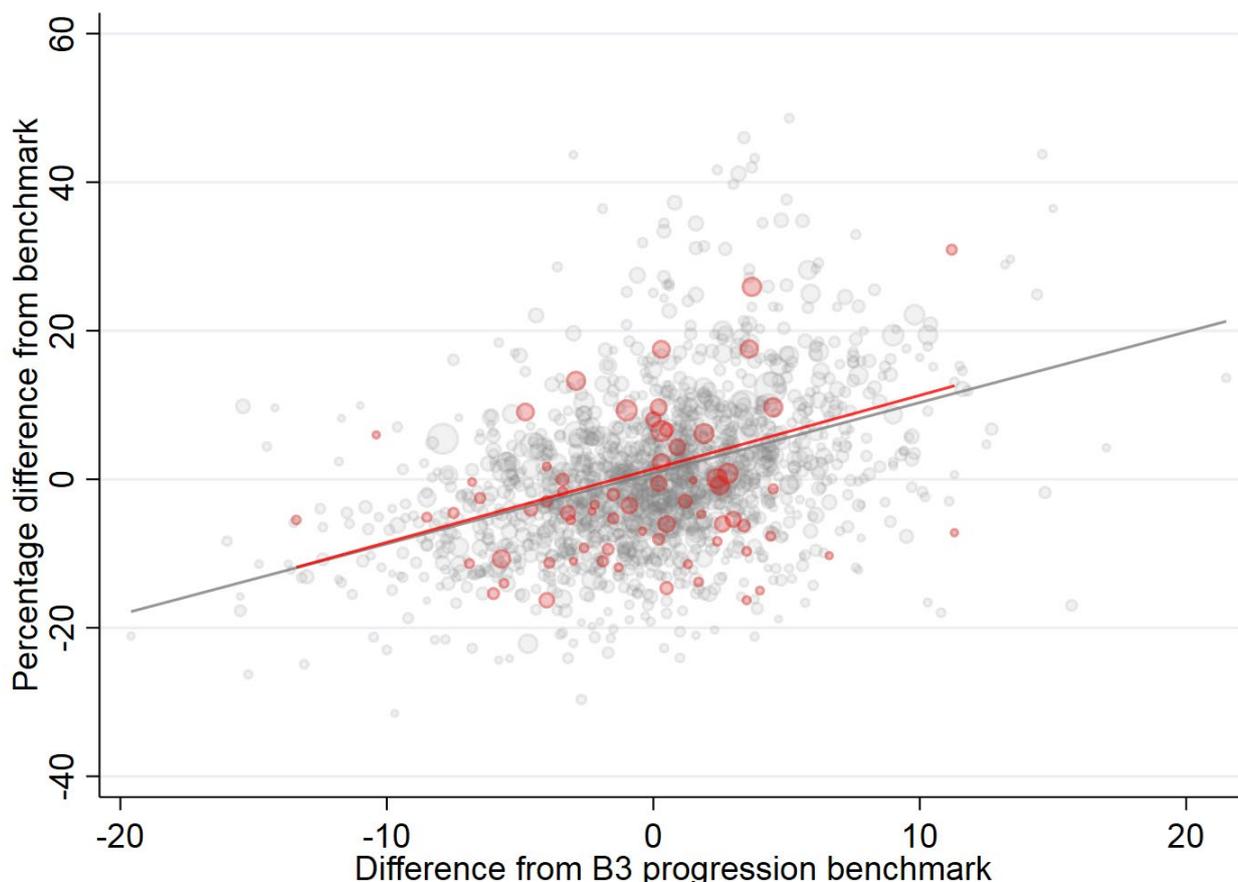
Figure 18 demonstrates that while the existing OfS progression indicator and graduate earnings are strongly related, they measure distinct things. However, a large part of the relationship will come from picking up the same differences in labour market success by subject and student background. Indeed, Figure 26 in the appendix shows that the relationship between benchmark earnings and the benchmark progression indicator is just as strong as the relationship between actual earnings and the actual progression indicator value. A key question is whether the differences from the respective benchmarks are similarly correlated.

Figure 19 plots the two measures against each other. The size of each bubble again represents the number of FPE students on each course. History courses are again highlighted in red. The grey and red lines are linear fit lines for all courses and for history, respectively.

The overall correlation coefficient is 0.39 (weighted by FPE), while it is similar (0.35) for history. This correlation is far from perfect, with some courses that score well on the existing progression metric scoring much less well on the earnings metric and vice versa.

These differences could point to the value of having two somewhat different indicators of labour market success based on different cohorts. Some courses may be good at helping their students into professional or managerial employment, but less good at preparing them for high-earnings jobs – or vice versa. Similarly, because these estimates are based on different graduation cohorts, the lower correlation could be partially driven by a provider getting lucky with its graduates' labour market outcomes for one set of cohorts but less lucky with the next set of cohorts, for example. These considerations support the adoption of an earnings metric as a *complement* to the existing OfS progression metric, rather than a replacement.

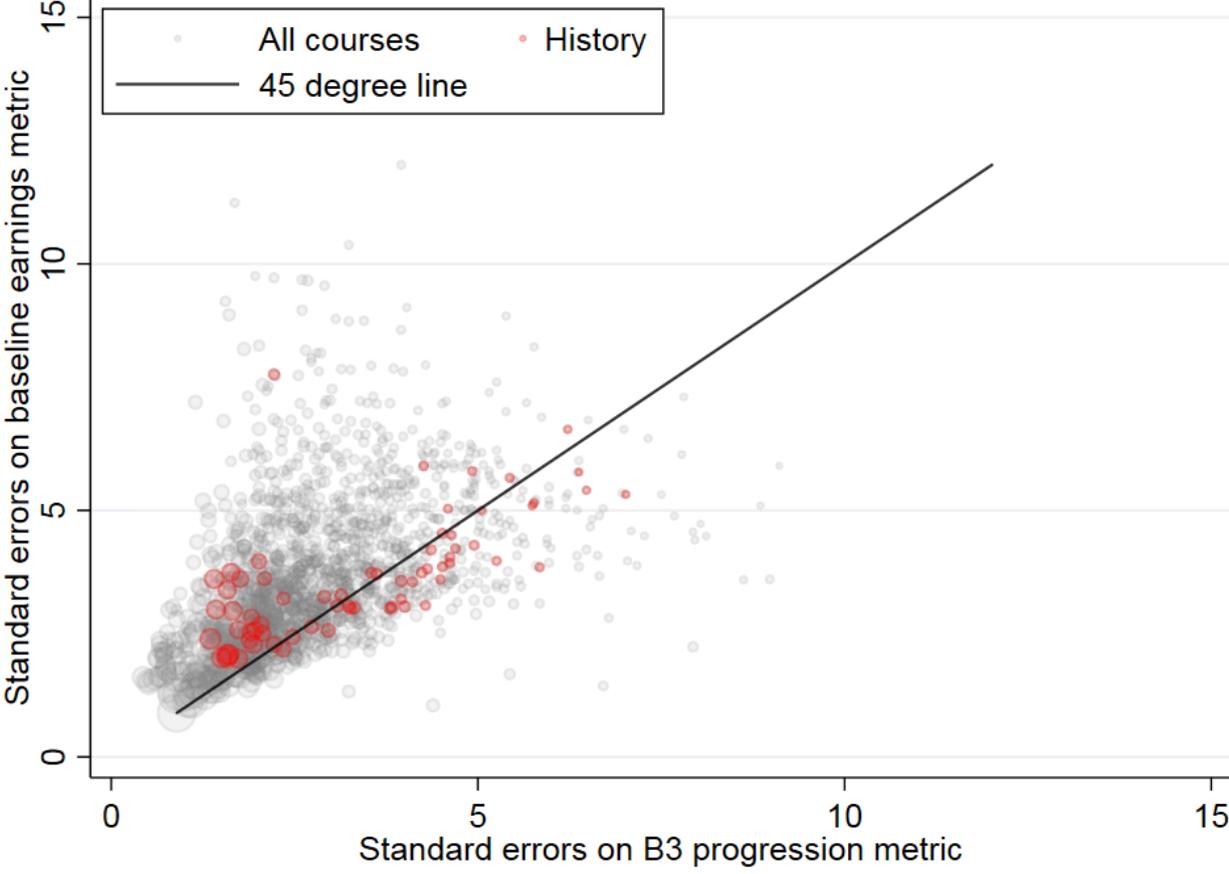
Figure 19: Difference from benchmarks for baseline earnings metric and B3 progression indicator, by course



Note: Plots the difference between a course's indicator value for the existing B3 progression metric (by subject for the 2017/18 to 2020/21 graduation cohorts) against our baseline estimate of the difference between actual average earnings on a course and benchmark earnings. Includes only full-time first degree courses where both estimates are available. The size of the bubbles reflects the geometric mean of the number of FPE students in the OfS sample and in our sample for each of the estimates. History courses are highlighted in red.

Finally, Figure 20 returns to the issue of **accuracy**, by comparing estimated standard errors for the B3 progression metric and the baseline earnings metric. While these standard errors are not strictly comparable – the B3 progression metric is an absolute share, whereas the baseline earnings metric is a percentage difference from an estimated benchmark – they nevertheless illustrate the level of statistical uncertainty attached to the two metrics. As expected, larger courses – as reflected by the larger bubble sizes in the plot – have smaller standard errors, and the size of the standard errors is well correlated between the two metrics. Overall, standard errors for the earnings metric are slightly but not dramatically larger than for the existing B3 progression metric.

Figure 20: Standard errors on baseline earnings metric and on B3 progression metric, by course



Note: Includes only full-time first degree courses where both estimates are available. The size of the bubbles reflects the geometric mean of the number of FPE students in the OfS sample and in our sample for each of the estimates. History courses are highlighted in red.

8. Sensitivity analysis

This section examines the sensitivity of our estimates to a range of alternative methodological choices. We compare our baseline estimates, as described in Sections 6 and 7, with results obtained from slightly different specifications. By examining these variations, we aim to provide a transparent view of how methodological decisions influence our findings and to demonstrate the reliability of our core results across different analytical approaches. Our focus remains on full-time first degree courses throughout this sensitivity analysis, though the same methodological variations could also be applied to other level–modes.

We assess how different specifications affect key properties of our earnings metric, including:

1. **Consistency in identifying course performance:** the extent to which courses' differences from benchmarks remain consistent across specifications. To capture this, we introduce:
 - *Metric A:* correlation with the baseline. Technically, this is the correlation of the within-bucket ranks across the baseline and alternative specifications.
 - *Metric B:* average absolute change in the course effect (as measured by the percentage difference from benchmark) between the baseline and an alternative specification, weighted by the number of students on each course under the baseline specification.
2. **Consistency in identifying underperforming courses:** whether courses flagged as having below-benchmark earnings remain consistent across specifications. For this, we introduce:
 - *Metric C:* proportion of courses where the course effect is significantly negative (using 95% confidence intervals) in the baseline which no longer meet this condition under the alternative specification.
3. **Selectivity relationship:** how the difference from benchmark correlates with course selectivity. For this, we introduce:
 - *Metric D:* correlation between courses' difference from benchmark and their selectivity (measured by average total GCSE points of graduates). This is calculated within subject and is a simple average across subjects.

In each subsection that follows, we present an alternative specification and compare its results with our baseline. We focus on how these key metrics change, highlighting specific points of interest. The tables are colour-coded to indicate which variations of our baseline specification make an important difference, as outlined by the table below. This

colour-coding is based on our own subjective assessment and is therefore only included as a guide for the reader.

Colour-coding for sensitivity checks

Baseline model
No important difference relative to the baseline
Noteworthy difference relative to the baseline
Important difference relative to the baseline

In the appendix, we provide a set of plots of the course effects for full-time first degrees in history to help visualise the impacts of the various sensitivity checks (Figures 28–34). Finally, we note that Metrics A–D listed above are not the only sensitivity checks we consider. In fact, we subject these different variations to a much wider battery of tests which are too extensive to fully document here. We provide a comprehensive list of metrics in the appendix (Table 14) and results for each specification in the accompanying spreadsheet.

i. Pooling across cohorts

In our baseline model, we pool across two graduation cohorts. Here, we consider the impact of pooling across a different number of cohorts. As discussed in Section 5.v, this trades off coverage and stability of estimates, against using data from further back in time which may make estimates less relevant and less responsive to changes in university performance.

We also compare estimates under the baseline specification for the same courses, where we estimate the model using earlier sets of graduation cohorts. This helps us to assess the stability of our estimates over time.

Our judgement is that increasing the number of cohorts included in the analysis from two to three does not make enough of a difference for full-time first degrees to justify doing it. Coverage would improve from 93% to 96%, and the median standard error would fall from 2.7 percentage points to 2.2ppts (both not reported in the table; the distribution of standard errors for different variations is shown in Figure 37 in the appendix). But changes in the other key metrics are relatively small.

Table 4: Different options for pooling across cohorts

	Metric A	Metric B	Metric C	Metric D	
Baseline				0.571	Pool the 2014/15 and 2015/16 graduation cohorts.
One cohort	0.934	0.020	0.324	0.551	Include only the 2015/16 graduation cohort.
Three cohorts	0.967	0.015	0.081	0.564	Include three cohorts of students (2013/14, 2014/15 and 2015/16).
Two years earlier	0.814	0.035	0.407	0.512	Estimate using data that are two years older than the baseline (for the 2012/13 and 2013/14 cohorts).
Four years earlier	0.812	0.037	0.342	0.512	Estimate using data that are four years older (for the 2010/11 and 2011/12 cohorts).

On average, course effects move by only 1.5ppts, while only 8.1% of courses identified as having significantly negative course effects in the baseline are no longer in this situation in the alternative.³⁸ The trade-off of estimating bigger models (which are more computationally demanding) and having to use data from graduates who attended university longer ago therefore does not seem worth it in this case. However, it is worth noting that increasing the number of cohorts in estimation might be advisable for the other level-modes which, as we have described above, suffer from small sample size issues.

However, based on the evidence here, we would argue against *reducing* the number of cohorts in estimation down to one. That is primarily because coverage would drop quite considerably (down to 84% from 93%) and standard errors would increase (from 2.7ppts in the baseline to 3.5ppts), but also because around one-third of the courses identified as significantly negative in the baseline would switch to not being significantly negative in the alternative. The fact that the overall correlation with the baseline remains high

³⁸ We consider this to be quite small. A large share of courses have estimates that are borderline significantly negative, so small changes in the model can result in apparently large changes in this metric. In practice, the regulator might choose a tighter threshold for regulation, such as an estimate significantly below 10% (see Section 10 for a discussion of this), so a significantly negative estimate would not actually be what is important for individual courses.

emphasises that this is largely driven by the reduction in precision (resulting in wider confidence intervals) due to the smaller sample sizes.

We see from the final two rows of the table that there are quite considerable differences in the baseline metrics over time. The overall rank correlation with the baseline is around 0.80 for the earlier cohorts, while the average change in the course effects is around 3.5ppts in their difference from benchmark, and more than one-third of courses identified as having significantly below-benchmark earnings would not have been two or four years earlier. Nevertheless, despite flagging this as yellow, we consider this performance over time to be quite good, as the average absolute change in the course effect is only slightly larger than the average standard error, suggesting that these differences are largely due to statistical uncertainty.

ii. Transformations of earnings outcome

Our baseline specification excludes those with very low earnings and Winsorises earnings at the top of the distribution. Here, we check the sensitivity of our estimates to those decisions, first by trimming at different points at the bottom of the distribution and second by Winsorising at different points.

Table 5 shows that small changes to the lower earnings threshold we use make only a minor difference. The only exception is that 18% of courses that have significantly negative returns in the baseline are no longer significantly negative when we reduce the lower threshold to £1,000. This is driven by more lower-earning graduates being included in the estimation of returns for already poor-performing courses, lowering the overall average. It also reflects that standard errors are typically higher when we reduce the lower threshold. This latter point makes us prefer the slightly higher threshold of £3,000, as the lower threshold adds noise, reducing precision.

When we increase the lower threshold up to the annualised national living wage, however, we do see more of a difference. As shown in Figure 35 in the appendix, which shows the full distribution of our earnings outcome variable, this removes 6.6% of graduates from the sample. This has little impact on coverage of full-time first degree courses, but would reduce the number of students on full-time Other UG courses we would report estimates for by around 10%. Around 12% of courses performing below benchmark performance in the baseline would not be picked up as doing so under this alternative. The correlation between difference from benchmark and selectivity would increase (from 0.571 to 0.638). As described in more detail in Section 9, if a much higher earnings threshold is chosen, we would recommend further analysis that considers course-level shares of graduates achieving earnings above the threshold.

Table 5: Different options for transforming the earnings outcome

	Metric A	Metric B	Metric C	Metric D	
Baseline				0.571	Drop earnings below £3,000 (in 2021/22 prices). This affects around 5.4% of full-time first degree graduates (including 4.0% with zero earnings). Winsorise earnings at the 99 th percentile within bucket and sex.
Trim at £1,000	0.966	0.014	0.183	0.521	Exclude anyone whose highest earnings three–five years after graduation are below £1,000.
Trim at £5,000	0.984	0.009	0.077	0.586	Exclude anyone whose highest earnings three–five years after graduation are below £5,000.
Trim at NLW	0.923	0.021	0.122	0.638	Exclude anyone whose highest earnings three–five years after graduation are below the annualised national living wage (which we set at £13,900 in 2021/22, assuming 30-hour weeks for 52 weeks).
Winsorise at £245,000	0.999	0.003	0.008	0.579	Cap maximum earnings at £245,000, rather than the 99 th percentile.
Winsorise at £100,000	0.999	0.003	0.005	0.576	Cap maximum earnings at £100,000, rather than the 99 th percentile.

Finally, although we think it is important to Winsorise in order to reduce the effect of really large outliers, we see that in practice the choice of how exactly we do this (within reasonable ranges) makes very little difference.³⁹

³⁹ We check sensitivity to Winsorising at £245,000 because this is the level above which HESA removes salary outliers from published Graduate Outcomes survey results, although concerns about data quality may be more acute with survey responses than with administrative data. See <https://www.hesa.ac.uk/support/definitions/graduates>.

iii. Specific earnings outcome

Table 6: Different options for the earnings outcome used

	Metric A	Metric B	Metric C	Metric D	
Baseline				0.571	Outcome is a graduate's earnings in their highest-earning year three–five YAG, adjusting for CPI.
Highest earnings four–five YAG	0.969	0.013	0.136	0.588	Use earnings in an individual's highest-earning year four–five YAG.
Earnings five YAG	0.945	0.020	0.214	0.595	Use an individual's earnings in their fifth full tax year after graduation. For the 2015/16 graduation cohort, this means using only earnings in the 2021/22 tax year.
Average earnings three–five YAG	0.972	0.014	0.150	0.509	Instead of using the highest-earning year for an individual, take the average of their taxable earnings across the three years.
Excluding earnings close to specific tax thresholds	0.999	0.002	0.005	0.571	Exclude recorded earnings that are up to £20 below specific thresholds in each tax year at which we observe bunching in the data. ⁴⁰
Excluding years when studying	0.980	0.012	0.088	0.625	Exclude earnings in any tax year in which an individual is observed to be studying.

The baseline specification uses a graduate's highest reported earnings in the period three–five years after graduation (YAG) as the key outcome variable. Here we investigate the impact of different choices of outcome variable – over different numbers of years, or using an average instead of the highest-earning year. We also consider the impact of excluding earnings observations close to specific tax thresholds and in tax years when

⁴⁰ Specifically, the primary threshold for employee National Insurance contributions, the secondary threshold for employer NICs and the personal allowance.

individuals are observed to be studying, as in both cases recorded earnings may be a poor reflection of a graduate's earnings potential.

Most of these alternative choices for the outcome variable make little difference to our estimates. All are very highly correlated with the baseline, with rank correlations of at least 0.94, while the average change in point estimates is no more than 2ppts.

We do see some changes in which courses would be identified as having significantly negative point estimates. Around 14% of courses performing below benchmark in the baseline would not be picked up as doing so if we maximised earnings over two years rather than three, and around 21% would not be if we used only one year of earnings. We have seen already that this measure can be quite noisy, and sensitive to changes in the sample (and the sample does get slightly smaller when we reduce the number of years of data we use, due to more people earning below the minimum earnings threshold).

We also see that 15% of courses are no longer significantly negative when we switch to average earnings over three–five YAG rather than highest earnings. While our preference is to take the highest earnings over this period because we suspect average earnings may be affected by periods when graduates are not fully engaged in the labour market, we nevertheless consider using average earnings a reasonable alternative.

Finally, we consider dropping those in further study. We find that this makes little difference in practice, so we do not condition on this measure in our baseline estimates in the interest of simplicity. The only metric that moves appreciably is the correlation with selectivity, which strengthens because it is those graduating from more selective courses who are more likely to undertake further study. One reason this makes little difference overall may be that many of those flagged as studying in a given tax year appear to have relatively high earnings. More than two-thirds of those in the 2015/16 graduation cohort flagged as studying in their fifth tax year after graduation earned above the annualised national living wage in that tax year (compared with 82% amongst those not flagged as studying). This may be because the measure includes any study during the tax year, and many are working alongside their studies or only study for a small part of the year.

iv. Controlling for prior attainment

Our baseline model aims to adjust fairly for differences in student characteristics across providers. In particular, we include in our model controls for students' prior attainment and their background characteristics. Here we consider a few variations on what we control for and the way some of these controls enter the model.

Table 7: Different options for prior attainment controls

	Metric A	Metric B	Metric C	Metric D	
Baseline				0.571	Includes controls for total GCSE (and equivalent) points, GCSE English points and GCSE maths points. Each is standardised within Year 11 cohorts. Full details can be found in Table 12 in the appendix.
More flexible KS4 attainment controls	0.999	0.002	0.005	0.578	Instead of standardised GCSE point scores (total, maths and English), include the raw scores and a square term in each. This is the simplest way of allowing for a non-linear relationship between prior attainment and earnings.
Add KS2 attainment controls	0.988	0.008	0.097	0.550	Include additional controls for English and maths KS2 point scores, standardised within KS2 cohorts.
Add KS5 attainment controls	0.999	0.003	0.023	0.558	Include additional controls for the total points from A level and equivalent qualifications, standardised within KS5 cohorts.
Attainment controls only	0.949	0.024	0.135	0.564	Only include controls for prior KS4 attainment and academic year in which they were in Year 11 (a rough proxy for age), but no other controls for student background.
Demog. controls only	0.979	0.022	0.038	0.688	Only include controls for student demographics, and not controls for prior attainment.

First, we include more flexible controls for GCSE ('KS4') attainment. The primary motivation for doing this is to explore whether it weakens the observed positive correlation between the course effect estimates and HEP selectivity. In practice, however, this makes very little difference: the correlation with the baseline is 0.999, while

the correlation with selectivity is barely different from that in the baseline (0.578 versus 0.571). We therefore prefer the simpler approach to controlling for GCSE scores.

Second, we consider the inclusion of controls for Key Stage 2 attainment. While this could improve our ability to control for student ability, there is a downside in that we lose some of the sample due to imperfect matching to the KS2 records. In practice, since we observe that this makes very little difference, we think it is preferable to omit KS2 scores from the model.

Third, we consider controlling for Key Stage 5 attainment. As described in Section 5.iv, this is challenging due to the wide variation in the types of things people study at KS5 level, and comes with a potentially additional challenge that KS5 grades are typically not finalised before university admission decisions are taken. We control in a simple way for KS5 scores here, finding that it does not make much difference. On balance, our preference is therefore to not do this. However, this is something that could potentially be explored in more detail in future – for example, by attempting to model the match between subjects studied in higher education and courses taken at KS5 level.

Finally, we consider only using a subset of the control variables – first, only attainment controls and, second, only demographic controls. Both these models produce results that are similar to the baseline, although there is some movement here, most notably in the strength of the relationship between course effects and selectivity in the second case. Nevertheless, we recommend including both sets of controls in the main model.

v. Location when earning

An important feature of our baseline model is that we *do not* control for where individuals live while they are working, even though this information is available in the tax records. However, since graduate earnings vary substantially across the country, one might argue that providers in some regions might be disadvantaged due to their lack of proximity to strong local labour markets.

Here we try a few different approaches to adjusting for where graduates live as working adults. First, we directly control for the region they are living in five years after graduation; second, we control for whether they are living in London (only); and finally, we adjust our outcome variable each tax year to reflect differences in living costs between areas.

Table 8: Different options for accounting for location when earning

	Metric A	Metric B	Metric C	Metric D	
Baseline				0.571	Include controls for home region (government office region when they applied to the course) and for quintile of area deprivation, but not for location after graduation.
Region when earning	0.943	0.021	0.224	0.522	Include additional control for the government office region they are living in (based on HMRC records) five YAG and exclude those where their location is Abroad or Unknown.
London TTWA when earning	0.969	0.015	0.159	0.528	Include additional control for whether they were living in the London or Slough and Heathrow travel-to-work areas five YAG and exclude those where their TTWA is unknown. This approximately captures those living within commuting distance of the London labour market.
Adjust earnings with local living costs	0.924	0.022	0.232	0.484	Adjust earnings each tax year to reflect approximate relative living costs in the local authority area in which a graduate was living in that tax year. This reflects median rents for a two-bedroom property in each area in 2021/22 relative to the median in England and assumes housing costs account for 24% of spending (the weight on housing in CPIH in 2021). This effectively discounts earnings of someone living in Islington, London by 25% and increases earnings for someone living in Barnsley, South Yorkshire by 11%.

We find that all three variations on the baseline in this category make a moderate difference to the estimates. In all three cases, the correlation between the metrics remains high (especially when just controlling for whether a graduate is living in London, where the correlation is 0.97), and we see fairly small average estimate changes of around 2ppts. However, 16–23% of courses switch from having significantly negative effects under the baseline to not being significantly negative under the alternative scenarios. And finally, the relationship between course effects and course selectivity weakens in all three cases, reflecting the fact that students from more selective universities are much more likely to move to high-earning cities, and especially London.⁴¹

Consistent with the discussion in Section 5.iv, we would not recommend that the model used to generate the earnings metric control for the region in which graduates are living as working adults. In short, our rationale is that location after leaving university is a ‘bad control’, so controlling for it would move the earnings metric away from measuring the (full) impact of a course on the earnings of its graduates. Specifically, it would shut down an important channel through which providers can have an impact on their graduates’ earnings – by enabling them to move to stronger labour markets. We note that controls for graduate’s home region and socio-economic status when applying to university already control for relevant geographical differences in providers’ *intake*.

However, we acknowledge that controlling for region when earning may nonetheless be attractive for policymakers seeking to avoid penalising providers serving local communities. Our suggestion is for the regulator to take into account additional contextual information, as discussed in Section 6. In addition, results with controls for region when working could be presented as supplementary ‘region-of-earning-adjusted’ benchmarks alongside the main estimates.

If the regulator determined that this was insufficient and was instead keen to proceed with an adjustment for where graduates live as working adults in the main earnings metric, our recommendation would be to use our third approach here of adjusting the outcome variable for regional price indices. This would account for the higher cost of living in strong labour markets, which is largely due to higher housing costs. As living costs and local labour market earnings are strongly correlated, the effect of adjusting for local living costs on the earnings metric would in practice be qualitatively similar: providers with many graduates working in London (i.e. highly selective and London-based providers) would lose out and others would gain.

We would prefer this adjustment over controlling for region when earning, as in contrast to adding region-when-earning controls, it would preserve the interpretation of the earnings metric as seeking to measure the impact a course has on its graduates’ (real) earnings. It is very plausible that, for instance, higher earnings in London for nursing and

⁴¹ Figure 32 in the appendix highlights this for history: the institutions that are most affected by the regional adjustments are institutions with high benchmarks, which are by and large the most selective (see Figure 10), and London-based institutions.

medicine graduates do not or only barely make up for higher living costs, while earnings are much higher for law graduates even when higher living costs are accounted for.⁴² In this case, adjusting for regional price levels would still give credit to law courses enabling graduates to move to London, while an earnings metric with controls for region when earning would miss this mechanism entirely.

Unfortunately, however, robust official regional price indices are currently unavailable. Our estimates here rely on imperfect rough regional price indices constructed from local housing costs (see Table 8 for details). Better data on regional price levels would make adjusting for local living costs much more compelling.

vi. Controlling for course-level characteristics

Here we consider different options for dealing with certain course-level characteristics. As discussed in Section 5.iii, this is impossible within our baseline ‘fixed effects’ approach; we therefore show results for a ‘mean-residual approach’ and a ‘hybrid random / fixed effects approach’. The primary focus here is on factors that do not vary within courses, such as its average selectivity (as measured by the average GCSE score of its graduates) or the geographical location of the university. We also consider the composition of subjects within the broader CAH2 subject classification.

The mean-residual and both hybrid approaches we present here all make a substantial difference to the course estimates compared with the baseline. Most notably, all three (by construction) completely remove the relationship between course selectivity and course returns, which is very strong in the baseline. They also (unsurprisingly) dramatically affect which institutions are identified as having an earnings metric that is significantly different from zero.

Whether to control for course-level selectivity is clearly a major decision for regulation. Given the strong positive relationship between selectivity and the baseline earnings metric, controls for course selectivity very substantially reduce earnings metric estimates for the most selective providers and increase them for less selective providers. The impact of an additional control for whether a HEP is located in London is comparatively minor.

For the reasons discussed in Section 5.iii, we would on balance recommend the fixed effects approach for the primary earnings metric without any course-level controls. However, we believe there is a strong case for supplementing this with a ‘selectivity-adjusted benchmark’ obtained using the hybrid approach with course-level selectivity controls. While based on stronger assumptions than the primary benchmark, we think this additional benchmark would provide important context for regulation.

⁴² See figure A11 in Britton et al. (2021) for suggestive evidence for this.

Table 9: Different options for accounting for course-level characteristics

	Metric A	Metric B	Metric C	Metric D	
Baseline				0.571	Fixed effects regression, with no course-level controls (which cannot be separately identified from HEP dummies). Comparisons are within broad subject area (CAH2 level).
Mean-residual approach, with controls for selectivity	0.753	0.049	0.573	−0.006	Mean-residual estimation approach, as described in Section 5.iii, with additional course-level controls for selectivity, measured as the average total, maths and English GCSE point scores amongst graduates.
Hybrid random / fixed effects approach, with controls for selectivity	0.754	0.049	0.537	−0.019	Hybrid random / fixed effects approach, as described in Section 5.iii, with course-level controls for selectivity.
Plus control for whether HEP in London	0.741	0.050	0.573	−0.016	Hybrid random / fixed effects approach, with controls for course selectivity and an additional control for whether the registered address of the HEP is in the London or Slough and Heathrow TTWAs.
Detailed subject area	0.968	0.015	0.173	0.586	Include controls for the detailed subject area of a course (at CAH3 level). For history and archaeology, this means including dummies for history; history of art, architecture and design; archaeology; heritage studies; and classics.

Finally, we attempt to address the fact that the subject definition we use is an aggregated measure of several more detailed subjects. The composition of these subjects may vary considerably across universities in a way that may harm some providers unfairly.

Although we do not consider it practical to estimate course effects at a more detailed subject level (due to small sample sizes and incompatibility with the OfS framework), we instead propose an alternative specification here which controls for subject composition within the CAH2 definitions. We find that this makes a moderate difference to the results, with 17% of courses moving from having significantly negative estimates under the baseline to not under the alternative.

The main downside of controlling for detailed subject area in this way is that the detailed subject coefficients and the provider effects are only separately identified if there are students taking different detailed subjects at the same provider. If, for example, heritage studies was the only history subject at providers teaching it, no earnings metric could be calculated for history courses at these providers. Even if this was only nearly the case, the earnings metrics for these providers could not be robustly estimated due to the collinearity between the provider dummy and the detailed subject dummy. In deciding whether to control for detailed subject, the regulator would therefore need to trade off greater fairness across providers against statistical robustness.

vii. Non-parametric approaches

Our baseline approach uses a parametric linear regression framework. Here we investigate alternative non-parametric approaches which instead group ‘like’ individuals who share many characteristics and compare the earnings of each graduate with average earnings of those in the same group. We divide individuals into just over 2,500 potential bins based on combinations of the following characteristics:

- male
- seven-point measure of socio-economic background in Year 11 (eligible for free school meals, attended independent school; otherwise, quintile of area deprivation)
- home region when applied (nine government office regions)
- academic year of graduation
- ten bands of GCSE attainment, based on percentiles of total GCSE points amongst whole Year 11 cohort.⁴³

This approach involves making an important trade-off between having sufficient bins for comparing ‘similar’ students, while still having few enough that the majority of students have a comparator. The approach means comparing the earnings of, for example, a male from the North East who was eligible for free school meals in Year 11, had total GCSE points between the 70th and 75th percentiles amongst the Year 11 cohort nationally and who graduated in 2014/15, with the earnings of others with that combination of characteristics.

⁴³ With cuts at the 50th, 60th, 70th, 75th, 80th, 85th, 90th, 95th and 98th percentiles.

Table 10: Non-parametric approaches

	Metric A	Metric B	Metric C	Metric D	
Baseline				0.571	
Non-parametric (mean-residual)	0.951	0.027	0.344	0.470	Non-parametric approach comparing earnings with those of others in the same 'bin' based on a range of characteristics. Implementation analogous to mean-residual approach.
Non-parametric (fixed effects)	0.971	0.015	0.048	0.582	Non-parametric approach with the same 'bins', but implemented in an analogous way to our baseline, fixed effects regression.

We present two possible non-parametric approaches here. In the first, which is similar to a mean-residual approach, we find moderate changes relative to the baseline. The rank correlation is high, at 0.95, while the average change in course effects is moderate at 2.7ppts. However, we do see a large share (34%) of the courses that had significantly negative estimates under the baseline switching to not having significantly negative estimates under the alternative. Finally, the selectivity relationship weakens, from 0.57 to 0.47. We would not recommend using this approach as it relies on similarly strong assumptions to a parametric mean-residual approach, as discussed in Section 5.iii.

Finally, the second variant we consider is more similar to a fixed effects approach, and we find that this makes only small differences relative to our baseline. We do not feel strongly about using the baseline method or a non-parametric approach along these lines for the earnings metric.

9. Labour market participation

A feature of the tax data is that many individuals record very low levels of earnings, suggesting that they are not participating in the labour market for much of the tax year. As discussed in Section 8, in our baseline specification, we exclude all graduates from the estimation who have recorded earnings of less than £3,000 per year in 2021/22 prices in all three tax years three–five years after graduation. As we have argued in Section 5, excluding very low earners from any earnings estimates is justified on the grounds that such earnings are very unlikely to reflect these graduates’ true earnings potential. Instead, they are likely to represent part-time earnings or graduates being out of the labour force for all or part of the year – for example, because of further study, caring responsibilities, emigration, incarceration, health problems or death.⁴⁴

However, excluding graduates with very low earnings also raises the concern that our estimates could be missing an important part of some courses’ causal effect on earnings. As an extreme example, a course might enable a small minority of graduates to have very high earnings but lead to very low earnings for a large majority. In that case, if all the low-earning graduates were dropped from the sample, the course might score very well on our proposed earnings metric even though it delivered very poor outcomes for most of its graduates.

This concern is more important with a higher trimming threshold. A relatively low threshold of £3,000 per year in our baseline approach only excludes 5.4% of graduates who would otherwise have been included (including 4.0% with zero earnings). A higher threshold – such as full-time earnings on the national living wage, which were £13,900 in 2021/22⁴⁵ – would exclude 12.9% of graduates.

In our view, there are three reasonable ways of addressing this concern:

- 1. The share of low-earning graduates excluded from the estimation due to low earnings could be displayed as an additional column on the B3 dashboard.**

This would be straightforward to implement and easy to interpret, and would provide important context for any earnings metric estimates. In the hypothetical case discussed above, it would be straightforward to see that the earnings metric estimate was based on a small minority of a course’s graduates.

However, merely displaying the ‘raw’ share of excluded graduates would provide no context for interpreting the share of low-earning graduates itself. It would not be clear whether this share was high or low compared with courses in the same subject area and given the characteristics of graduates of a given course. In our

⁴⁴ Those with *zero* earnings in all of the three tax years cannot be included in the model, as the natural logarithm of zero is undefined.

⁴⁵ The hourly national living wage was £8.91 for people aged 23 and over in 2021/22. We calculate a conservative annual full-time equivalent as $£8.91 \times 30$ (hours) $\times 52$ (weeks) = £13,900.

view, this would not be satisfactory if the share of low-earning graduates was substantial and varied a lot across subjects and with the characteristics of a course's intake.

- 2. Like (1), but in addition displaying a 'benchmark' share of excluded graduates as a separate column.** This would require estimating models of the same structure as for the primary earnings metric, but with the share of excluded graduates as the outcome variable. As for earnings, the course-level benchmark would be the predicted share of excluded graduates, leaving out the coefficient on the provider dummy (see Section 5.iii).

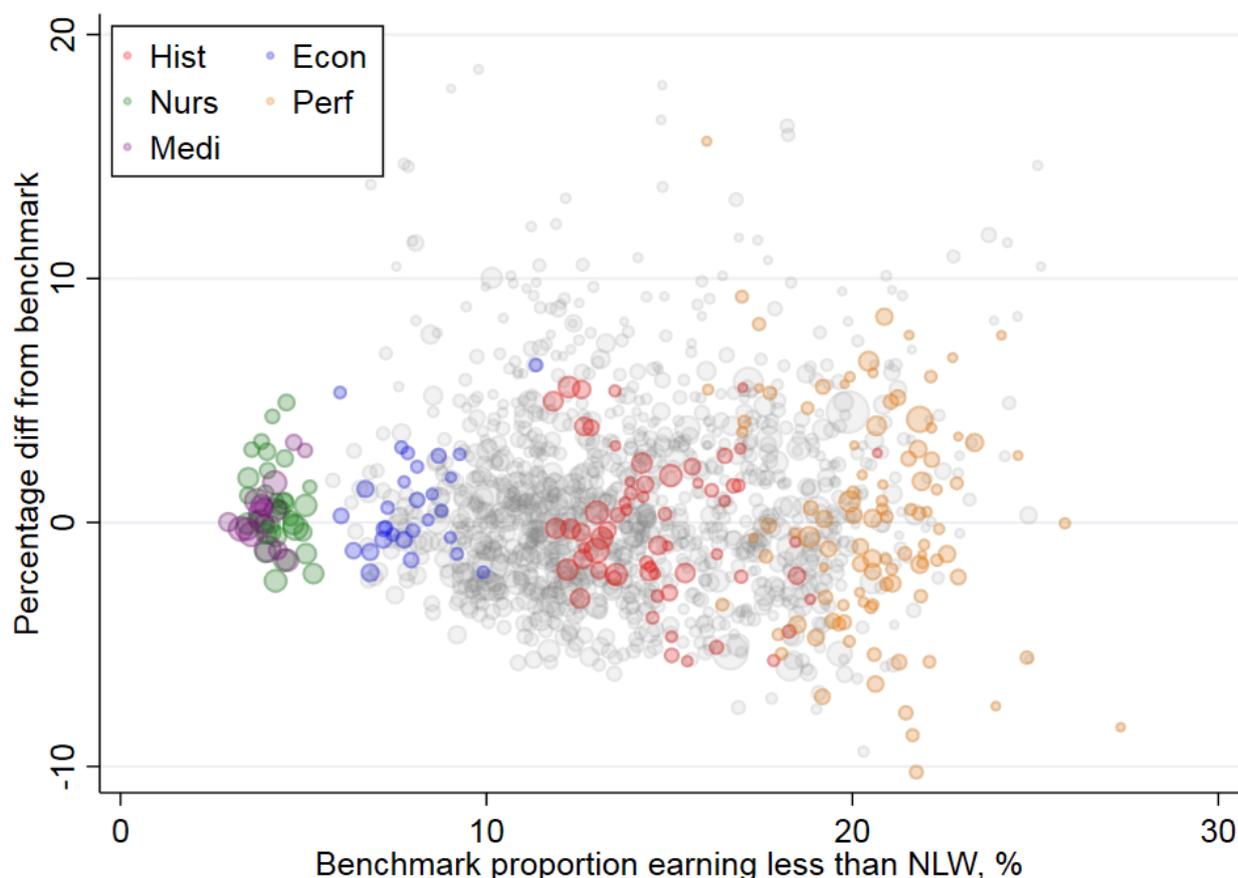
A benchmark share of excluded graduates would add the required context to the actual share, as for each course the share of low-earning graduates could be compared with the benchmark. However, this way of displaying the results would not give any indication of the statistical uncertainty to which these estimates would be subject. This is an important limitation: even with a relatively high exclusion threshold, the number of graduates with low earnings would be small for many courses, leading to substantial statistical uncertainty.

- 3. Like (2) but presenting the difference between the share of excluded graduates and the corresponding benchmark as a separate progression metric.** This third progression metric could be displayed on its own page in the B3 dashboard. As for the other B3 metrics, statistical uncertainty could then be presented in the form of confidence bands around the estimates.

This would only be appropriate if the share of excluded (or included) graduates in fact picked up differences between courses in preparing their graduates for the labour market. This would be the case if graduates earning below a threshold did so because, for example, they could only find precarious employment in their chosen profession. It would not be the case if earnings below the threshold reflected other factors such as further study, caring responsibilities, etc.

We would recommend the third option if a relatively high threshold such as full-time earnings on the national living wage (NLW) was chosen. Figure 21 shows the benchmark for the proportion of graduates earning less than the NLW compared with the percentage difference from the benchmark. There is huge variation in the benchmark across courses between around 3% for some medicine courses and more than 25% for some performing arts courses. Overall, the variance of the benchmark accounts for 63% of the variance in the share of students earning less than the NLW. (Note that the chart excludes the 18% of courses teaching around 9% of students for which we provide estimates for our primary earnings metric, but which have fewer than 10 FPE graduates in the sample earning less than full-time earnings on the NLW in all three relevant tax years.)

Figure 21: Benchmark share earning less than the NLW against percentage difference from benchmark



Note: Includes all courses passing our minimum sample size restrictions and the statistical disclosure restriction that percentages need to relate to at least 10 students. The size of each bubble represents the number of FPE students on each course. History courses are highlighted in red, economics courses in blue, nursing courses in green, performing arts courses in orange and medicine courses in purple. Figures relating to 18% of courses (9% of students) that we would report estimates for are not shown due to statistical disclosure control.

Notably, the ordering of benchmarks for the highlighted subjects is very similar (albeit inverted) to the ordering for our main earnings benchmarks, with low benchmark shares for medicine and economics and high shares for performing arts courses. This strongly suggests that the proportion of graduates earning less than the NLW is picking up genuine differences in the labour market experience of different courses' graduates. (Consistent with its strong performance on the existing B3 progression metric, nursing stands out for the exceptionally low share of its graduates earning below the NLW despite its middling earnings outcomes overall.)

Figure 22 shows the same picture for the proportion of graduates earning less than £3,000 per year. For nearly all courses, the benchmark share is below 10% (58% of courses teaching 35% of all graduates are excluded from the chart because fewer than 10 FPE of their graduates fell below the threshold). Less than half of the variance in the

share earning less than £3,000 is accounted for by the variance of the benchmark, as shown in Table 15 in the appendix. While relative differences in benchmark between courses are still substantial, percentage point differences are minor. The benchmark is also much less strongly related to subject choice, making it plausible that the differences mostly reflect factors not directly related to graduates' labour market experience such as further study, caring responsibilities, etc. We therefore conclude that adding the share of graduates earning less than £3,000 as an additional column (option 1 above) would be the best course of action if our baseline metric was adopted.

Figure 22: Benchmark share earning less than £3,000 per year against percentage difference from benchmark



Note: Includes all courses passing our minimum sample size restrictions and the statistical disclosure restriction that percentages need to relate to at least 10 students. The size of each bubble represents the number of FPE students on each course. History courses are highlighted in red, economics courses in blue, nursing courses in green, performing arts courses in orange and medicine courses in purple. Figures relating to 58% of courses (35% of students) that we would report estimates for are not shown due to statistical disclosure control.

10. Recommendations

This section distils the findings of this report into recommendations for the government to consider. Our recommendations range from the most general – whether the Office for Students should consider using administrative earnings data at all – to the specific – for example, how many cohorts should be pooled for estimating earnings benchmarks. In many cases, there is a range of reasonable options, which we attempt to map out as far as possible.

The following list sets out our recommendations in words. Table 11 presents all the options we considered in tabular form, with a verdict and brief justification for each recommendation.

- **We recommend adopting an earnings metric calculated from administrative data in addition to the existing B3 progression metric.** An earnings metric would complement the existing B3 progression metric, revealing a more holistic picture of providers' success, or otherwise, in equipping graduates with labour market skills. Specifically:
 - As shown in Section 7, earnings capture a related but different facet of labour market success from the 'positive outcome' measured by the B3 progression metric.
 - The earnings metric we propose would measure labour market success somewhat later in graduates' careers, when outcomes will likely be more predictive of lifetime outcomes.
 - Using administrative data would provide evidence on those graduates (more than half of the total) who do not respond to the Graduate Outcomes survey.
 - An earnings metric would, at any point in time, relate to earlier cohorts than the B3 progression indicator and thus provide statistically independent evidence of a provider's performance.
- **While the benefits of adopting an earnings metric need to be weighed against some drawbacks, on balance we think the advantages predominate.** In our view, appropriate contextualisation of the result can largely avoid negative effects such as incentivising an undue focus on labour market skills at the expense of other aspects of course quality, or penalising providers for enabling graduates to obtain jobs in their chosen field where these are lower paid. We also judge that the cost of calculating an additional metric for the OfS would be modest, as a similar metric has previously been calculated for the TEF (the cost for providers would be minimal, as no additional data would need to be submitted). However, a full cost–benefit assessment is beyond the scope of this report.

- **In contrast to the existing B3 progression metric (but in line with the progression outcome for the TEF), we recommend using the difference from a course-specific benchmark as the headline measure.** As we have argued in Sections 3 and 5.i, there are compelling reasons to focus on a benchmark targeting the causal effect of a course on its graduates' earnings relative to other courses in the same subject area. We therefore recommend changing the default visualisation on the B3 dashboard for an earnings metric to 'Difference from benchmark values only'.
- **We recommend presenting actual average earnings for a provider alongside differences from the benchmark, as is currently done in the 'Difference from benchmark values only' view of the B3 dashboard.** This would provide useful context, allowing stakeholders to distinguish, for example, cases where the earnings metric was negative and actual average earnings were low from those where the earnings metric was negative but actual earnings outcomes were still high. Given the potential for misinterpretation, we do not recommend presenting actual average earnings in comparison with a regulatory threshold that is the same for all providers.
- **A numerical threshold for the difference from benchmark could be added and reflected in regulation.** A threshold of -20% for first degree undergraduates would leave courses teaching 0.8% of all graduates below the threshold. A threshold of -10% for first degree undergraduates would leave courses teaching 9.4% of graduates below the threshold (1.7% significantly so). However, we believe that adding an earnings metric to the OfS's B3 indicators could provide helpful context for regulation even if no numerical threshold was defined.
- **We recommend that the underlying benchmarking model for an earnings metric control for prior attainment and a range of demographic characteristics.** This would be consistent with the OfS's current approach to benchmarking for the existing B3 metrics. We take no view on whether this model should be parametric or non-parametric.
- **We prefer estimation approaches such as 'fixed effects' that require no assumption about the correlation between the earnings metric and individual-level control variables.** However, we can also see advantages of other approaches such as the current OfS approach to benchmarking for the B3 metrics. In the choice of model and estimation approach, we recommend that a large weight be placed on transparency.
- **We recommend that the outcome variable used be based on earnings in tax years between two and six years after graduation.** Using the highest earnings three to five years after graduation provides a reasonable compromise in our view, but many variations are arguable. We recommend using the natural logarithm of this value as the outcome variable, as is standard in labour economics.

- **We recommend pooling data across two to three cohorts of graduates to maximise statistical power.** Most of the benefits of pooling arise with only two cohorts, but pooling three cohorts would also be reasonable. There is a strong case for working with cohorts that do not overlap the cohorts used for the existing progression metric, to capture fully independent information.
- **We do not recommend controlling for the location of people when earning.** The location choices of graduates are affected by their higher education experience, so controlling for location when earning would move the earnings metric away from measuring the (full) impact of a course on the earnings of its graduates. Instead, we recommend controlling for graduates' home region and reporting contextual information about graduates' location decisions separately, such as the share of graduates staying in the same local area as the provider or the share moving to London.
- **However, we acknowledge that policymakers may still wish to make an adjustment reflecting where graduates live as working adults to avoid penalising providers serving local communities.** A better alternative to adding controls for location when earning in our view would be adjusting earnings for regional price levels, but this is difficult to do in practice. The case for such a price-level adjustment would be substantially strengthened if robust regional price indices became available.
- **Calculating a 'selectivity-adjusted benchmark' controlling for course-level selectivity would provide helpful additional context.** This could be shown as an extra column on the B3 dashboard. However, given the strong assumptions required, we do not recommend for this to be the primary measure. We do not recommend controlling for provider location for this adjusted benchmark, because that would make the adjusted benchmark more difficult to interpret and less robust; it would also make little difference to the estimates in practice.
- **We recommend trimming the earnings outcome variable at the bottom and Winsorising it at the top.** Trimming at the bottom is advisable as some graduates may have unusually low earnings due to working part-time or taking time out of the labour market, so their current earnings may not reflect true earnings potential. Winsorising at the top of the earnings distribution can guard against extreme outliers having undue influence on individual courses' estimates. Our baseline approach is to trim annual earnings at £3,000 in 2021/22 prices and Winsorise them at the 99th percentile within level–mode–subject–sex combination, but many other approaches are arguable.
- **Showing the share of graduates earning below the trimming threshold provides helpful context for the interpretation of the earnings metric.** If a relatively high threshold such as full-time earnings on the national living wage, was chosen as the trimming threshold, we would recommend estimating separate models for the share of graduates earning below (or above) that threshold, which

could be adopted as a third progression indicator. If a relatively low threshold was chosen, as in our baseline approach, a separate column in the dashboard showing the share of graduates below the threshold would in our view be sufficient.

- **We would advise caution in reporting any earnings metrics for part-time and Other UG courses.** Estimates for these level–mode combinations will be substantially less robust than those for first degrees, partly due to the much lower number of students taught, and partly due to the larger share of mature students for whom school records are unavailable. In our view, there is therefore a good case not to proceed with earnings metrics for these level–mode combinations at least until match rates to school records improve. The OfS might also consider merging the first degree and UG with PG levels for the purposes of the B3 metrics.
- **We recommend updating any earnings metric estimates regularly as new data become available.** For the cohorts affected by teacher-assessed exams during the COVID-19 pandemic, it would in our view be reasonable to simply replace exam grades with teacher-assessed grades (provided that standardised GCSE scores are used). We would advise reviewing the methodological choice for the earnings metric whenever there are substantial changes to the available data in the future.
- **We recommend treating earnings metric estimates as one of many sources of evidence on provider performance.** No potential earnings metric would fully account for the complexity of the higher education landscape. Earnings are a more meaningful measure of labour market success in subjects such as business and management, where career success is quite closely linked to higher earnings, than in subjects such as nursing and medicine, where graduates are typically on fixed pay scales so pay differentiation is minimal, or in creative and performing arts subjects, where success as an artist may even entail lower earnings than an alternative career. It would therefore not be reasonable to treat poor performance on an earnings metric as an automatic trigger for regulatory action. As set out in the OfS’s regulatory framework, the wider context will need to be considered in each individual case.

Supplementing this list of recommendations, Table 11 sets out all the options we have considered for this report in tabular form, along with a brief justification for our recommendation. The goal is to give policymakers an overview of the decisions that our analysis leaves open, as well as the policy options that we believe should be ruled out. The rows are colour-coded according to the following key:

Colour-coding for policy options

This is our baseline recommendation.
This is a perfectly viable and reasonable alternative to our baseline recommendation.
This is not our main recommendation, but we think this choice would be arguable.
We do not recommend this option, having considered it empirically.
We reject this option for conceptual or practical reasons.

Except for the first category, the categories of options follow the order of the analytical decisions discussed in Section 5. The boxed options indicate our baseline decisions.

Table 11: Comprehensive table of options considered

Policy option	Brief justification for our recommendation
Category: level–modes to create a metric for	
Full-time first degree	<i>We believe an earnings metric for this level–mode category is viable.</i>
Full-time other undergraduate degree	<i>We are concerned about small/unrepresentative samples and advise caution in reporting earnings metrics for courses within this level–mode.</i>
Full-time undergraduate with elements of postgraduate	<i>We believe an earnings metric for this level–mode category is viable, but ideally it would be merged with full-time first degrees.</i>
Part-time first degree	<i>We are concerned about small/unrepresentative samples and advise caution in reporting earnings metrics for courses within this level–mode.</i>

Part-time other degree	<i>We are concerned about small/unrepresentative samples and advise caution in reporting earnings metrics for courses within this level–mode.</i>
------------------------	---

Category: counterfactual estimates are relative to

Estimates are relative to those who did not attend university	<i>It is difficult to adjust for selection into higher education with the controls at our disposal. We also think this places too much emphasis on factors that are beyond the control of institutions.</i>
Estimates are relative to the average student amongst everyone who went to university	<i>The main reason for rejecting this option is that it is difficult to adjust for selection into different subjects. We discuss several other reasons in Section 5.i.</i>
Estimates are relative to the average student amongst those who study a <i>similar</i> subject	<i>The same point as above applies, though to a lesser extent. This also does not align well with the existing regulatory framework.</i>
Estimates are relative to the average student within the same subject area (as defined by CAH2 classification)	<i>We think it is possible to adjust reasonably for selection into different institutions, conditional on subject area.</i>

Category: labour market outcome measure

- Timing

Maximum earnings three to five years after graduation	<i>This provides a balance between allowing graduates long enough to get established in the labour market and not waiting so long that that we are looking at graduates who attended the provider a long time ago. Taking the maximum over a few years smooths out some of the randomness in people’s earlier careers and is likely to be more predictive of longer-run outcomes.</i>
Earnings at a much earlier point (less than two years after graduation)	<i>This is not long enough for graduates to become established in the labour market.</i>
Earnings at a much later point (more than six years after graduation)	<i>This is too disconnected from universities’ current activities.</i>

Highest earnings four to five years after graduation	<i>This is very similar to taking the maximum over three years.</i>
Earnings five years after graduation	<i>Just taking earnings observed after five years amplifies the impact of random noise, creating more uncertainty in the estimates.</i>
Average earnings three to five years after graduation	<i>Taking the average gives very similar results to taking the maximum.</i>
- Sample selection	
Excluding earnings close to specific tax thresholds	<i>Although there is some bunching near thresholds, which could in theory drive bias, this is empirically unimportant.</i>
Excluding those who are studying	<i>Although excluding people in education is sensible, it does not matter much empirically and is more demanding in terms of data set-up.</i>
- Trimming and Winsorising	
Trim the earnings distribution at £3,000, excluding those below that level	<i>We exclude those with maximum earnings below £3,000 because people who are below this level are unlikely to be fully engaged in the labour market.</i>
Trim at £1,000	<i>Choosing £1,000 instead of £3,000 makes only a small difference but appears to add noise.</i>
Trim at £5,000	<i>Choosing £5,000 instead of £3,000 makes a negligible difference.</i>
Trim at approximately the annualised national living wage (£13,900 in 2021/22)	<i>This excludes many individuals. If this is chosen as the point to trim earnings, we recommend adding the share of individuals earning less than the threshold as an additional progression indicator.</i>
Winsorise earnings at the 99 th percentile	<i>We top-code to reduce the impact of outliers.</i>
Winsorise at £245,000	<i>Choosing £245k instead of the 99th percentile makes a negligible difference.</i>

Winsorise at £100,000	<i>Choosing £100k instead of the 99th percentile makes a negligible difference.</i>
- Adjusting the outcome variable	
Adjust earnings for living costs	<i>Adjusting the outcome variable for living costs in the area people live when they are working does affect the estimates. Our central recommendation is to not do this, largely because of imperfect data on living costs. With better data, this would be a very reasonable choice.</i>
- Alternative outcomes	
Share with earnings above a given earnings threshold	<i>If the NLW was chosen as the minimum level of earnings, we would recommend including this share as an additional progression indicator. With a lower threshold (say, £3k), we do not think this is necessary.</i>
Category: methodology	
Ordinary Least Squares with HEP fixed effects	<i>In the absence of more credible designs for identifying causal average treatment effects, this is the state-of-the-art in the academic literature.</i>
Non-parametric (mean-residual)	<i>We do not recommend this approach as it requires stronger assumptions. The main advantage of this type of approach would be the opportunity to control for provider selectivity (see discussion of controlling for provider selectivity below).</i>
Non-parametric (fixed effects)	<i>This produces very similar results to our baseline methodology.</i>
Category: control variables	
GCSE scores	<i>We recommend that an earnings metric adjust for GCSE scores.</i>
More flexible controls for GCSE scores	<i>We find that higher-order polynomial controls for GCSE attainment make very little difference.</i>

Student demographics (ethnicity / English as an additional language / free school meals etc.)	<i>We recommend that an earnings metric adjust for student demographics.</i>
KS2 attainment	<i>Controlling for this only makes a small difference to the estimates, but sample sizes are lower, as it is not possible to match to KS2 scores for all graduates.</i>
KS5 attainment	<i>We advise against controlling for this because it is difficult to control for KS5 attainment in a convincing way. A bespoke approach that controlled differently for KS5 attainment in different buckets might be appropriate. We are also concerned about KS5 results often being achieved after university admissions rounds.</i>
More detailed subject area controls	<i>These do make a moderate difference to the estimates and controlling for them is perhaps 'fairer', but it can cause problems with estimation in some cases.</i>
Region when earning	<i>We do not recommend controlling for graduates' location when earning, as enabling geographic mobility is one channel through which a course may impact graduates' earnings. Controlling for this would therefore move the earnings metric away from measuring the impact of a course on the earnings of its graduates. This option may nonetheless be attractive for policymakers seeking to avoid penalising providers serving local communities.</i>
London TTWA when earning	<i>As above.</i>
Selectivity (via mean-residual or hybrid random / fixed effects approach)	<i>We believe these are credible alternatives to our baseline approach. However, due to the strong assumptions required, our recommendation is instead to use one of these approaches to generate an additional 'selectivity-adjusted benchmark' to be reported alongside our main benchmark, which we believe would provide useful context.</i>

Selectivity <i>and</i> whether HEP is in London (via mean-residual or hybrid random / fixed effects approach)	<i>The estimates here are very similar to the selectivity-adjusted estimates discussed above.</i>
Use of machine learning techniques for selection of control variables	<i>We think this would be computationally burdensome and not very transparent.</i>
Set of higher education courses each student applied for	<i>Controlling for these is common in the academic economics literature and we would have liked to consider it, but we did not have the necessary data available.</i>
Category: sample size restrictions and cohort pooling	
- Sample size restrictions	
Minimum 50 individuals on a given course <i>and</i> the provider cannot account for more than 20% of the level–mode–subject’s students	<i>We believe it is important to only assess courses with reasonably large numbers of students. We also do not want any one course to be overly influential in a given bucket.</i>
Smaller or larger minimum sample restrictions to those above	<i>We did not re-estimate models with alternative sample size restrictions, but these could be considered. Small changes to the thresholds used here are unlikely to matter very much.</i>
- Pooling across cohorts	
Pool across two graduation cohorts	<i>We think this provides the best balance between getting large enough samples and timeliness of the metric.</i>
No pooling across cohorts	<i>This would reduce sample sizes, adding noise to the estimates.</i>
Pool across three cohorts	<i>This would mean going back further in time, which would reduce the timeliness of the metric.</i>

In addition to these options, there remain important policy decisions on how any earnings metric would best be *used*. These decisions are beyond the scope of this report. We make no recommendation on whether a numerical minimum threshold should be used as for the other B3 metrics and, if so, what it should be. We have no view on what the regulatory consequences should be of a point estimate below a numerical threshold, if such a threshold was defined, or how statistical uncertainty should be accounted for in

the decision about whether to take action or which action to take. We also do not take a stance on what other contextual data the OfS should consider before taking action – for example, whether it should only take action if actual average earnings of graduates were below a certain numerical threshold.

However, to inform future policy decisions, we do present statistics in Tables 16 and 17 in the appendix for various options on:

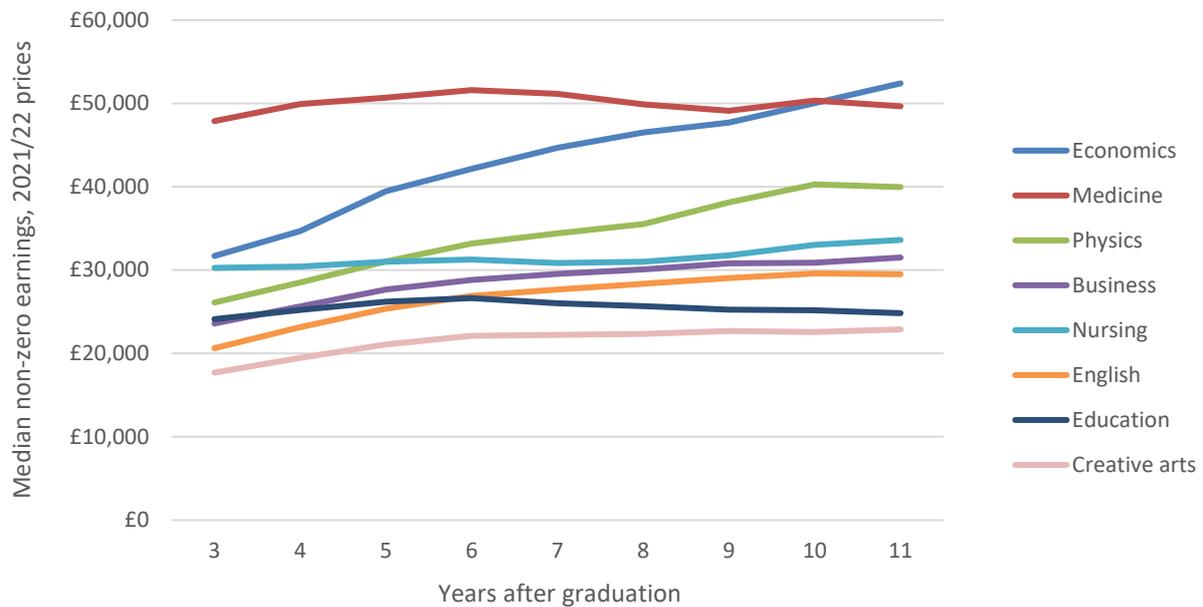
- i. the number of courses that would be ‘flagged’ for regulatory action;
- ii. the share of all students who graduate from ‘flagged’ courses;
- iii. the share of FSM-eligible students who graduate from ‘flagged’ courses;
- iv. the share of students at London-based providers who graduate from ‘flagged’ courses;
- v. the share of students on courses below the existing B3 progression threshold who graduate from ‘flagged’ courses;
- vi. the selectivity (compared with other courses in the same ‘buckets’) of the ‘flagged’ courses.

For each option, we also show in Table 18 how many courses in different broad subject areas would be affected.

In addition, Figure 38 in the appendix plots the proportion of graduates from courses that would be flagged as being below different numerical thresholds under our baseline approach. ‘Difference from benchmark’ shows the result if the threshold was simply applied to the point estimates and ‘Require significance’ shows what would happen if the whole 95% confidence interval was required to be below the threshold. This allows the reader to determine the consequences of different thresholds for the share of graduates affected by different numerical thresholds.

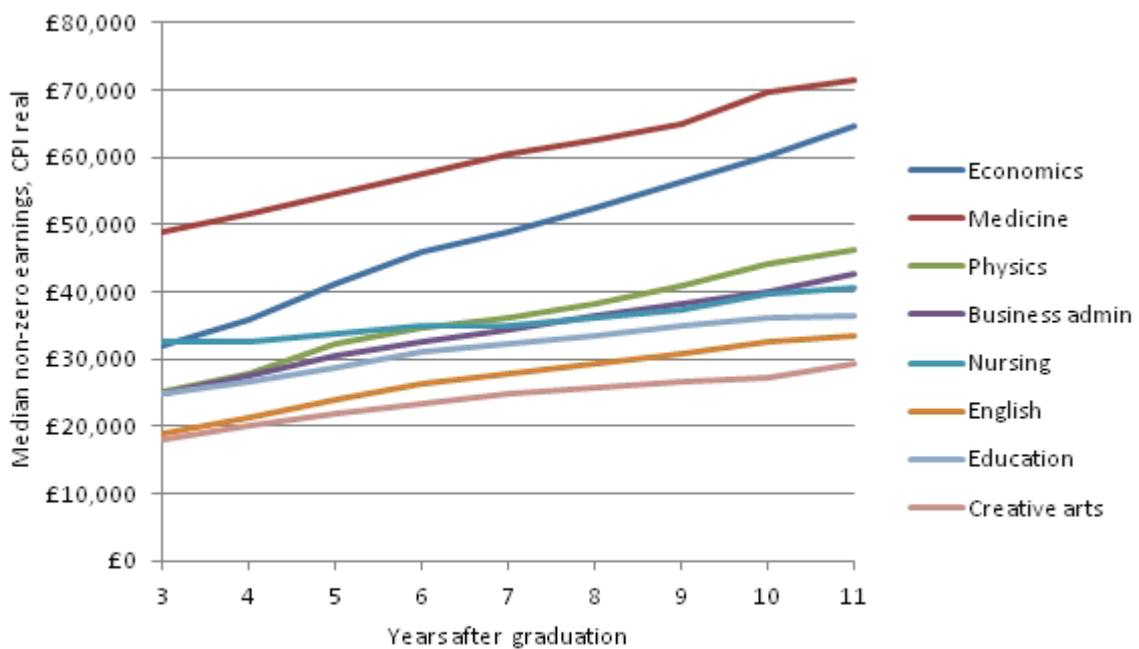
Appendix

Figure 23: Median earnings by subject area and years after graduation: women



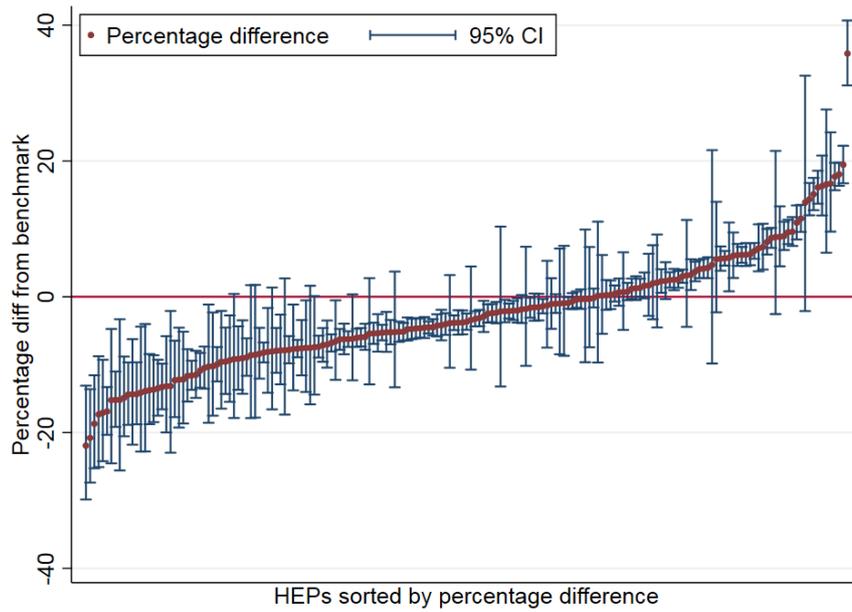
Note: See note to Figure 4.

Figure 24: Median earnings by subject area and years after graduation: men



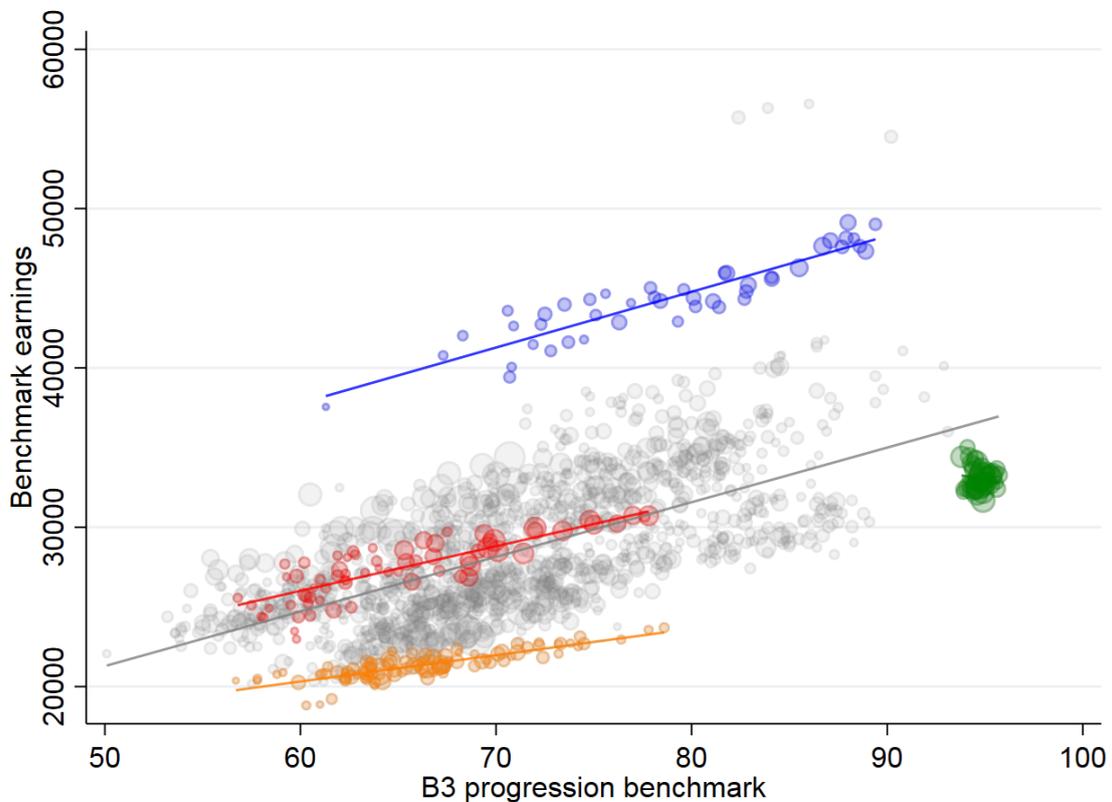
Note: See note to Figure 4.

Figure 25: Provider-level differences from benchmark, full-time first degree



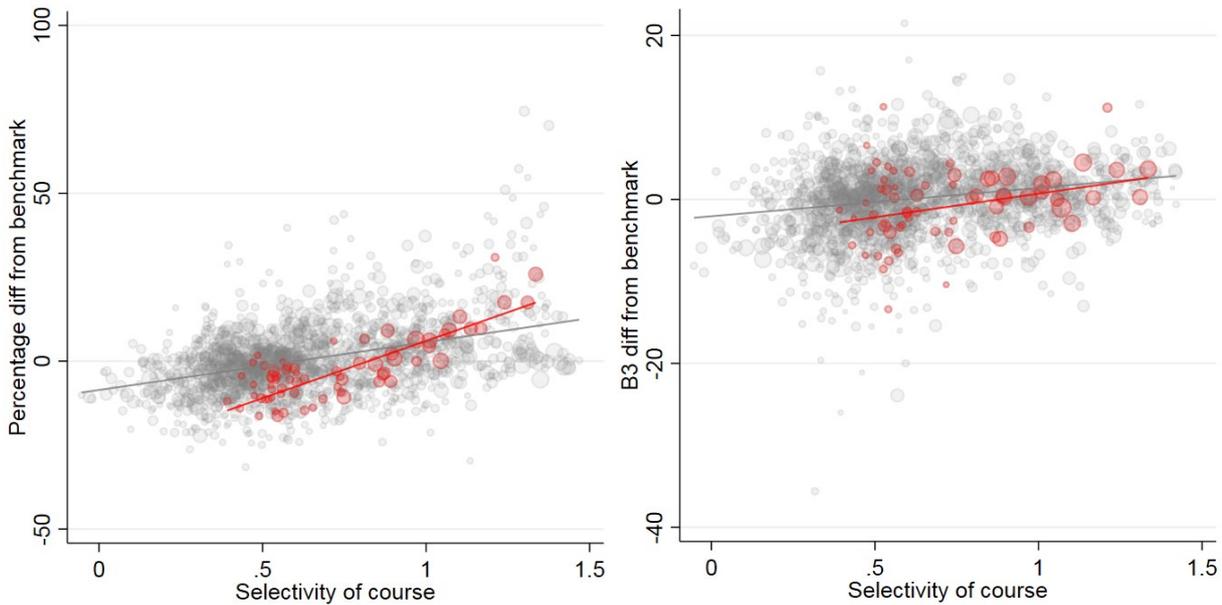
Note: Provider-level estimates for full-time first degree courses for all providers and subjects. Whiskers indicate 95% confidence intervals.

Figure 26: Benchmark earnings and OfS progression benchmark, by course



Note: Includes all courses for which we and the OfS both report estimates. The size of each bubble represents the geometric mean of the number of FPE students on each course across the OfS data and our data. History courses are highlighted in red, economics courses in blue, nursing courses in green and performing arts courses in orange.

**Figure 27: Percentage difference from benchmark and selectivity of course:
(a) baseline earnings metric and (b) OfS progression indicator**



Note: Includes all full-time first degree courses for which estimates are available. The size of the bubbles reflects the geometric mean of the number of full-person equivalent students in the OfS sample and in our sample for each of the estimates. History courses are highlighted in red.

Table 12: Control variables included in the baseline model

Variable	Description	Values	Mean
Male	Sex of the student, from HESA and ILR data (sex). Excluding those where sex is not recorded as male or female.		0.4380
Free school meals	Eligible for free school meals on census day in Year 11, from NPD school census (fsmeligible). Set to 0 for those attending independent schools.		0.0741
Independent school	Attended an independent school in Year 11, from NPD KS4 attainment records (toecode).		0.0792
English as an additional language	Whether a student is recorded as having English as an additional language, from the NPD school census (mothertongue, firstlanguage, languagegroupmajor). Set to 0 for those attending independent schools.		0.1924

Variable	Description	Values	Mean
SEN – statement	Pupil recorded as having a special educational need in Year 11, with a statement or EHCP, from NPD school census (senstage, senstatus, senprovision). Set to 0 for those attending independent schools.		0.0065
SEN – non-statement	Pupil recorded as having a special educational need in Year 11, with provision other than a statement or EHCP, from NPD school census (senstage, senstatus, senprovision). Set to 0 for those attending independent schools.		0.0788
GCSE total points	Capped total points from GCSEs and equivalent qualifications, standardised within Year 11 cohorts (including non-graduates) and Winsorised at 1% and 99% amongst graduates. From NPD KS4 attainment data (ptscoldg, ptscnewe, ptscneweptq, ptscneweptqee).		0.6053
GCSE maths points	Point score associated with grade achieved in full GCSE maths, standardised within Year 11 cohorts (including non-graduates) and Winsorised at 1% and 99% amongst graduates. From NPD KS4 attainment data (gcsemat, apmat, apmatptq, apmatptqee, apmat91).		0.5821
GCSE English points	Point score associated with highest grade achieved in full GCSE English, standardised within Year 11 cohorts (including non-graduates) and Winsorised at 1% and 99% amongst graduates. From NPD KS4 attainment data (gcseeng, apeng, apengptq, apengptqee, apeng91).		0.5681
Home region	Government office region in England at time of application, from HESA and ILR	North East	0.0507
		North West	0.1442

Variable	Description	Values	Mean
	(home_region). Excluding those with home region missing or null, or outside of England.	Yorkshire and the Humber	0.0984
		East Midlands	0.0852
		West Midlands	0.1087
		East of England	0.1091
		London	0.1603
		South East	0.1573
		South West	0.0862
Academic year in which Year 11	Academic year in which the pupil appears in NPD KS4 attainment data in Year 11 (yeargrp). Dummies for all years are included separately, but are grouped here for ease of presentation.	2001/02 to 2004/05	0.0644
		2005/06	0.0277
		2006/07	0.0443
		2007/08	0.0888
		2008/09	0.1967
		2009/10	0.3416
		2010/11	0.2153
		2011/12 or later	0.0211
Ethnic group	Broad ethnic group (self-reported) from HESA and ILR data. Unknown and missing values replaced with ethnic group as recorded in Year 11 from NPD school census.	Asian	0.1184
		Black	0.0540
		Mixed	0.0373
		Other	0.0116

Variable	Description	Values	Mean
		White	0.7787
Quintile of area deprivation	Quintile of percentile rank of IDACI score within Year 11 cohorts (including non-graduates), from NPD school census (idacis). This is a measure of small area deprivation, based on pupils' home postcodes. Unobserved for those attending independent schools.	Least deprived	0.2455
		2	0.2061
		3	0.1760
		4	0.1503
		Most deprived	0.1429

Note: Means are weighted by full-person equivalents and reflect the sample of students used in estimating the baseline model, from the 2014/15 and 2015/16 graduation cohorts. This is not restricted to graduates from full-time first degree students, but includes all buckets (combinations of level, mode of study and subject) for which we report any estimates. It includes those on courses for which results would not be reported.

Table 13: Coverage with alternative sample restriction of 23 FPE students (all restricted to buckets for which we report at least one estimate)

	All	First degree FT	First degree PT	Other UG FT	Other UG PT	UG with PG FT
[1] Total no. of courses	8,108	2,739	1,014	2,808	1,231	316
[2] No. of courses for which we report estimates (% of [1])	3,322 (41%)	2,059 (75%)	167 (16%)	673 (24%)	243 (20%)	180 (57%)
[3] Total no. of FPE students	698,200	482,900	44,400	89,800	55,400	25,800
[4] No. of FPE students on courses for which we report estimates (% of [3])	612,700 (88%)	469,900 (97%)	23,900 (54%)	58,500 (65%)	36,800 (67%)	23,500 (91%)

Note: FPE stands for full-person equivalent (see 'Definitions' in Section 1).

Figure 28: Earnings metric for full-time first degree courses in history: different options for pooling across cohorts

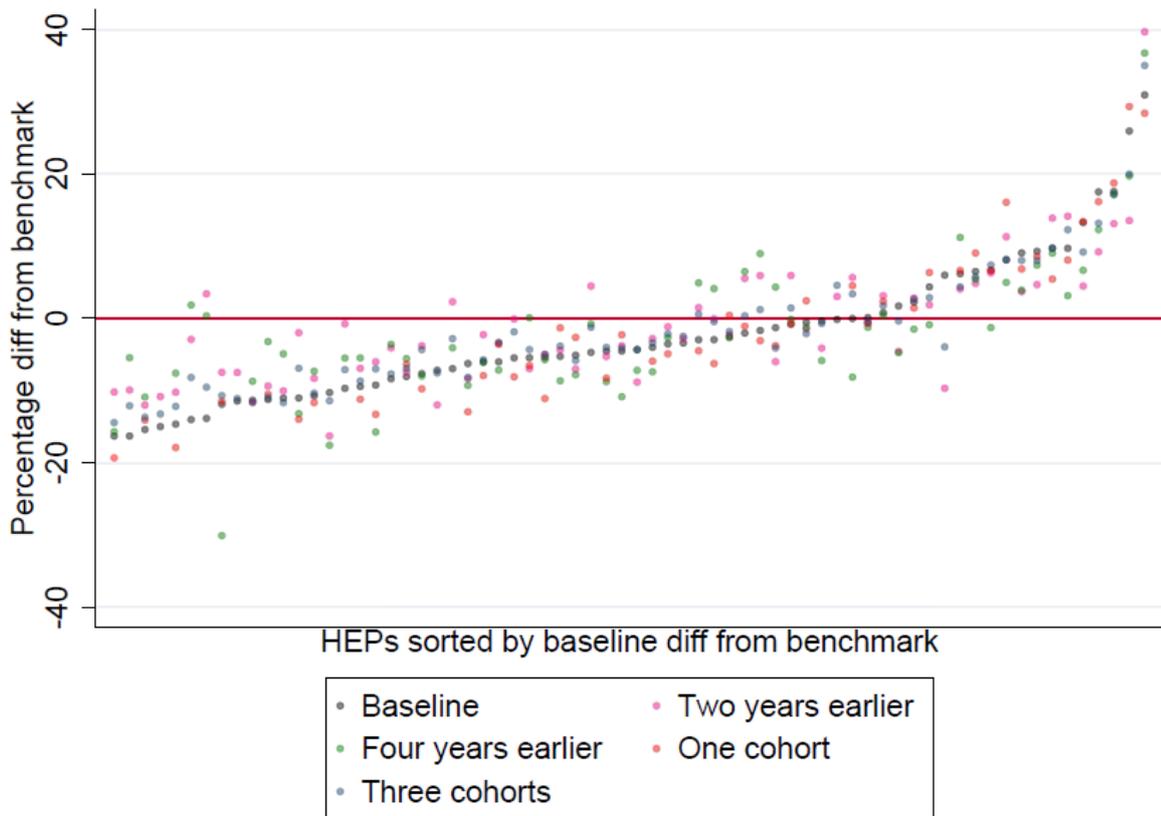


Figure 29: Earnings metric for full-time first degree courses in history: different options for transforming the earnings outcome

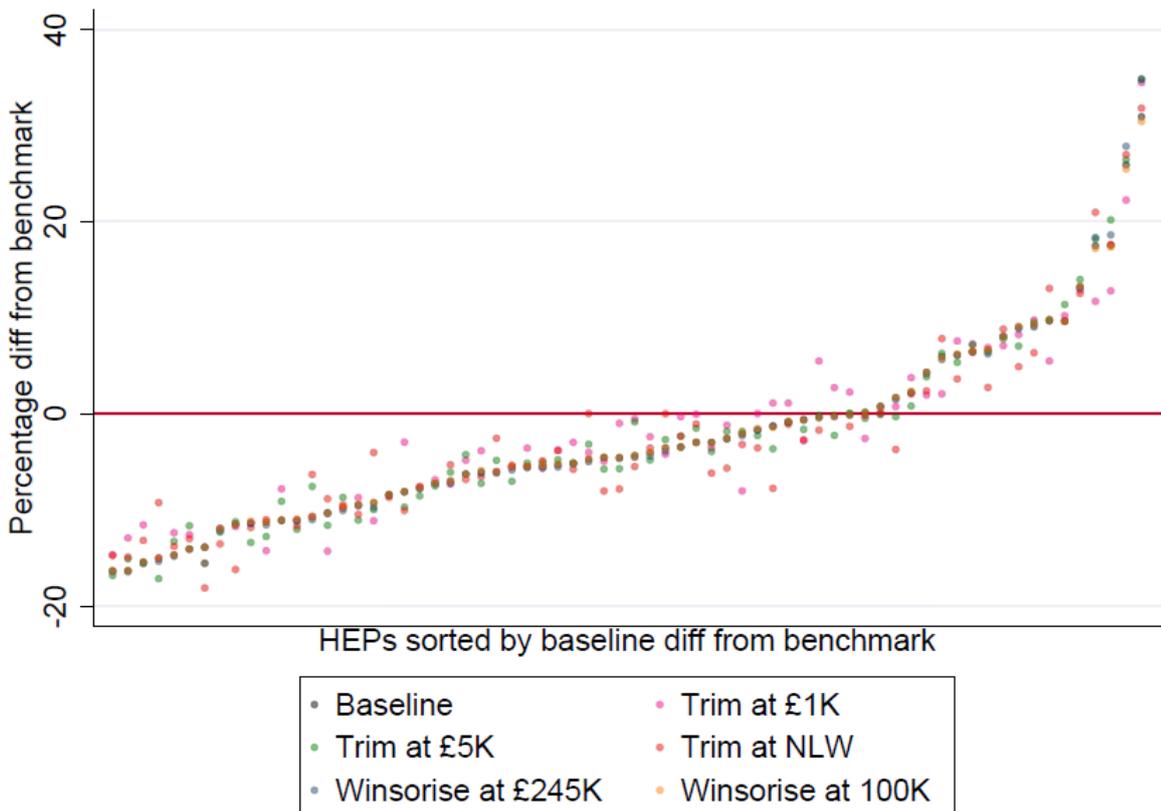


Figure 30: Earnings metric for full-time first degree courses in history: different options for the choice of earnings outcome

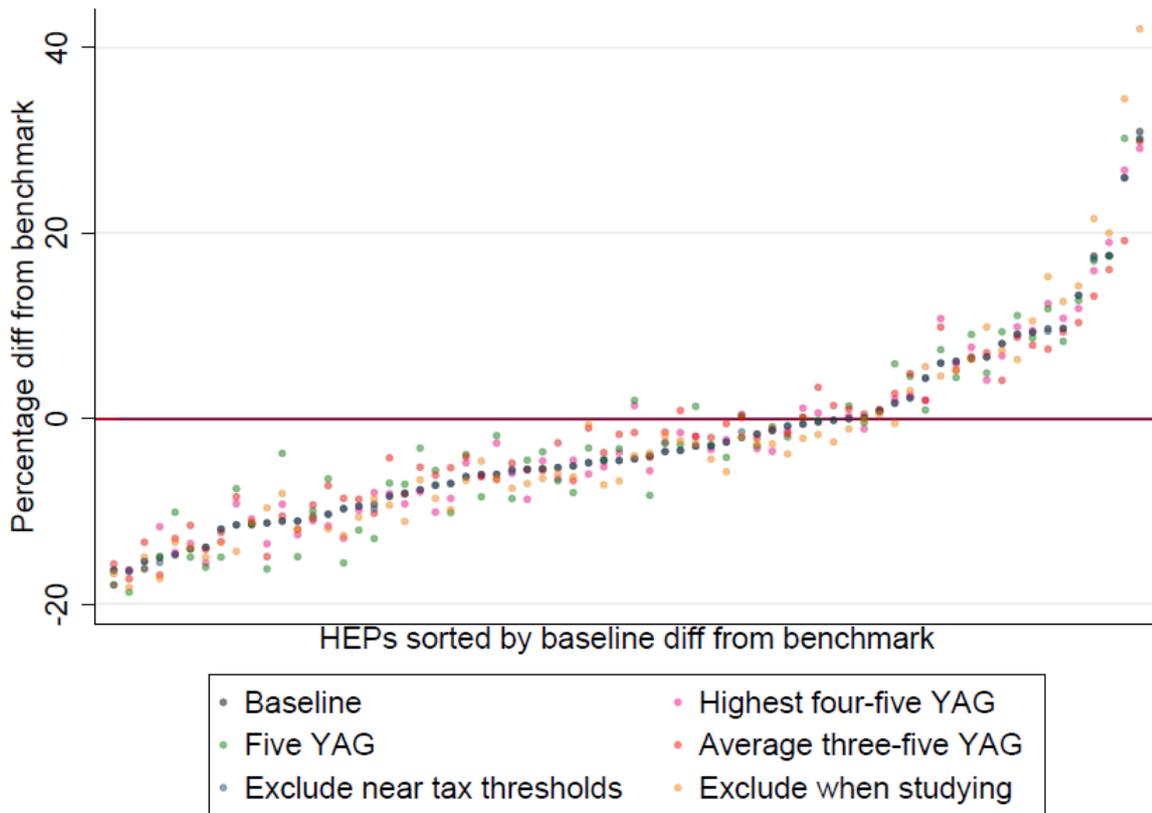


Figure 31: Earnings metric for full-time first degree courses in history: different options for controlling for prior attainment

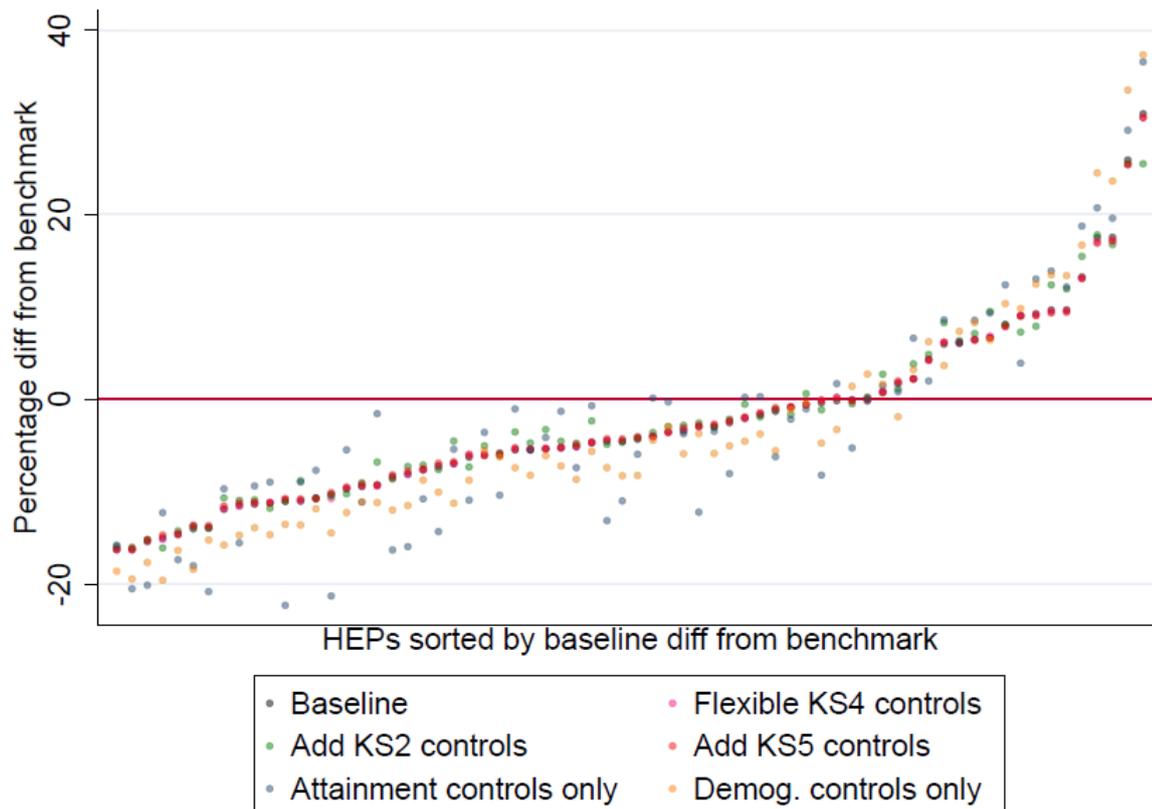


Figure 32: Earnings metric for full-time first degree courses in history: different options for accounting for location when earning

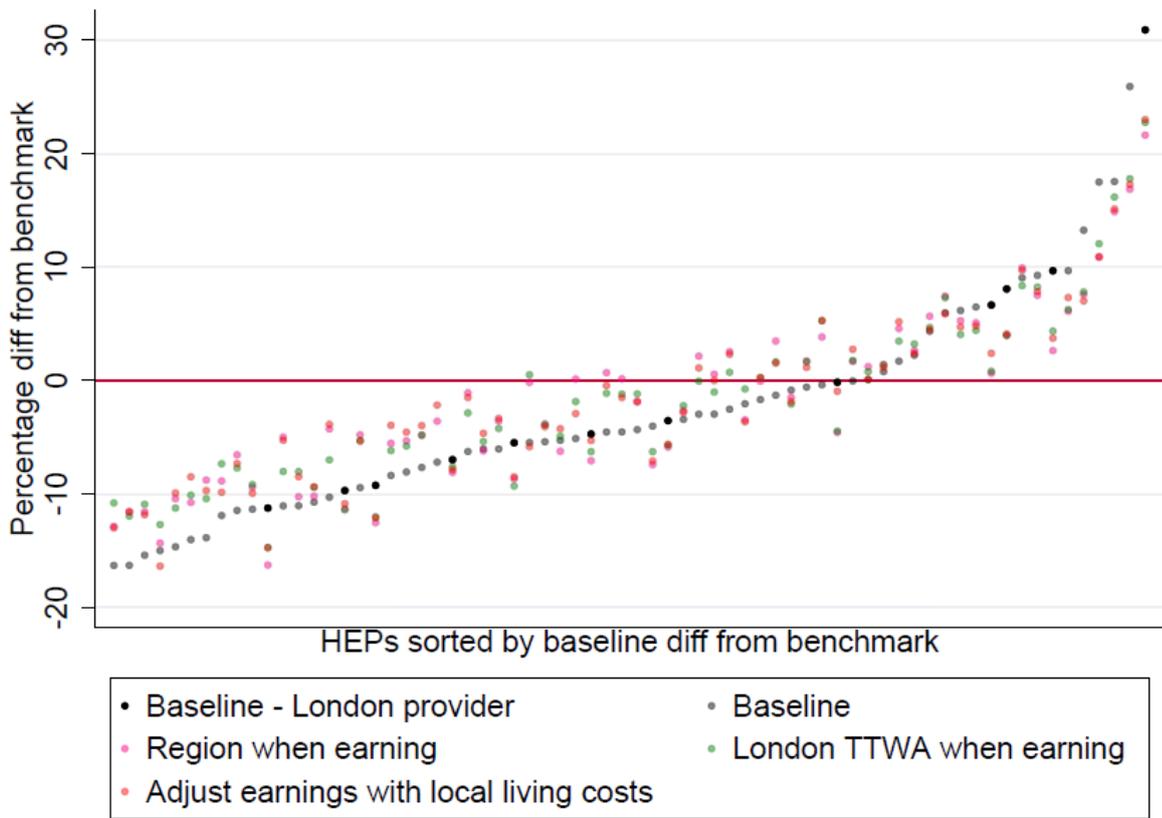


Figure 33: Earnings metric for full-time first degree courses in history: different options for controlling for course-level characteristics

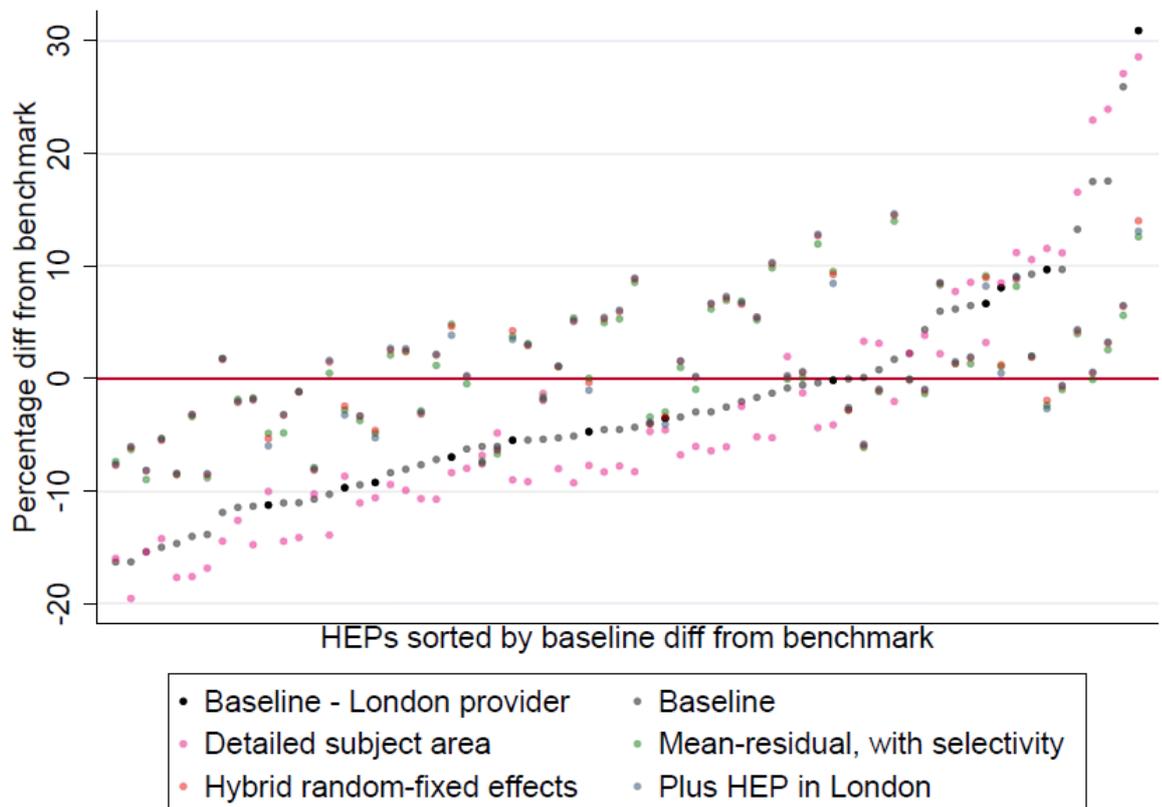


Figure 34: Earnings metric for full-time first degree courses in history: non-parametric approaches

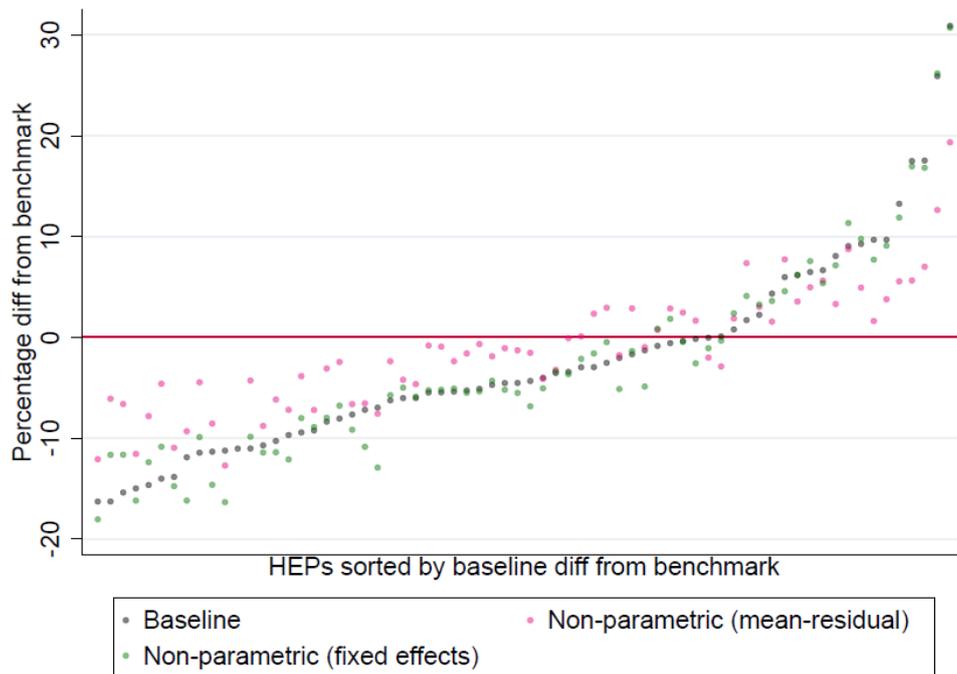
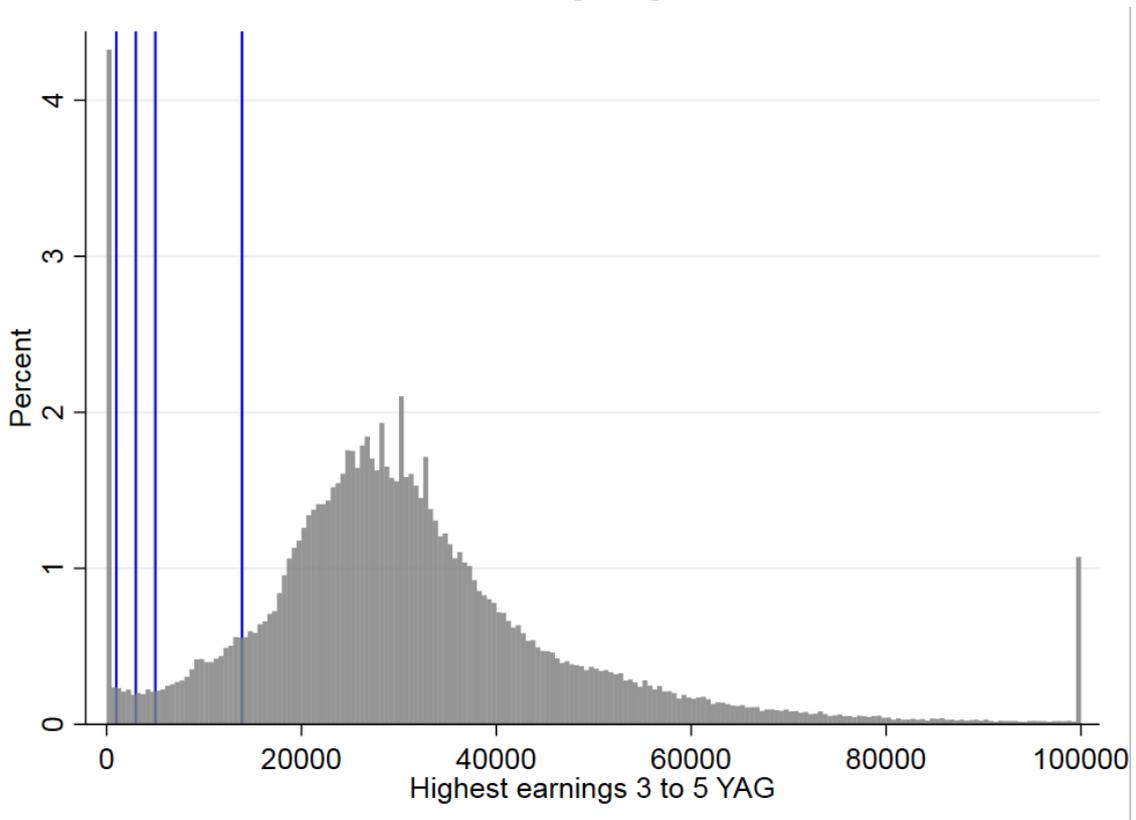


Figure 35: Distribution of highest earnings three–five years after graduation, full-time first degree graduates



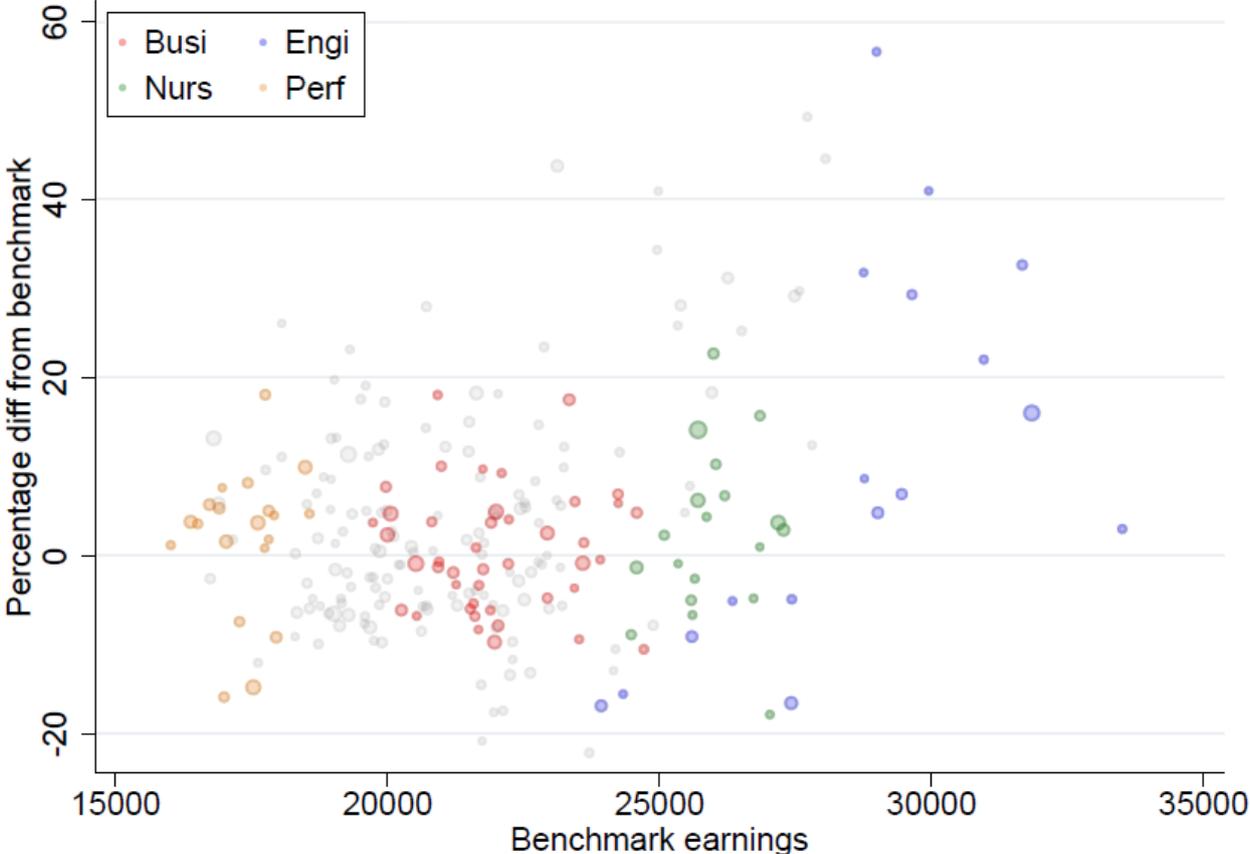
Note: Distribution of highest earnings three–five years after graduation for full-time first degree graduates from the 2015/16 graduation cohort. Winsorised at £100,000 in 2021/22 CPI real prices. Blue lines correspond to different lower thresholds discussed in Section 8.ii. These are £1,000, £3,000, £5,000 and £13,900.

Table 14: Full list of metrics used to examine methodological variations

Coverage
Proportion of courses for which we report estimates
Proportion of graduates who are from courses for which we report estimates
Proportion of graduates on courses for which we report estimates who are actually included in the estimation
Share of courses picked up as poor-performing
Proportion of courses for which estimated difference from benchmark is below –10ppt
Proportion of courses for which estimated difference from benchmark is below –20ppt
Proportion of courses for which estimated difference from benchmark is below –10ppt and actual average earnings are below £27,000
Proportion of courses where the difference from benchmark is significantly negative (using 95% confidence interval)
Relationship between benchmark and difference from benchmark
Proportion of courses where difference from benchmark is negative, which also have benchmark earnings in the bottom half within bucket
Proportion of courses where difference from benchmark is significantly negative (using 95% confidence interval), which also have benchmark earnings in bottom half within bucket
Proportion of courses where difference from benchmark is below –10ppt, which also have benchmark earnings in bottom half within bucket
Relationship between difference from benchmark and course selectivity
Correlation between courses' difference from benchmark and course selectivity (measured by average total GCSE points of graduates) (Metric D)
Correlation between percentile rank of difference from benchmark and percentile rank of course selectivity
Consistency in identifying course performance
Correlation between within-bucket ranks of difference from benchmark across specifications (Metric A)

Average absolute change in percentage point difference from benchmark between specifications, weighted by number of students on each course under the baseline specification (Metric B)
Root mean square deviation (RMSD) of within-subject percentile ranks of difference from benchmark between the baseline and alternative specification
RMSD of estimated difference from benchmark between the baseline and alternative specification
RMSD of benchmark earnings between baseline and alternative specification
Consistency in identifying specific courses as under- or over-performing
Proportion of courses where the course effect is significantly negative (using 95% confidence intervals) in the baseline which no longer meet this condition under the alternative specification (Metric C)
Proportion of courses where the course effect is significantly positive (using 95% confidence intervals) in the baseline which no longer meet this condition under the alternative specification
Proportion of courses with difference from benchmark below –10ppts and average earnings below £27,000 in the baseline which no longer meet these conditions under the alternative specification
Proportion of courses ranked in the bottom quintile of difference from benchmark within bucket in the baseline specification which are not ranked in the bottom quintile under the alternative specification
Proportion of courses ranked in the bottom quintile of difference from benchmark within bucket in the baseline specification which are not ranked in the bottom half under the alternative specification

Figure 36: Baseline benchmark and difference from benchmark estimates, all courses, full-time other undergraduates



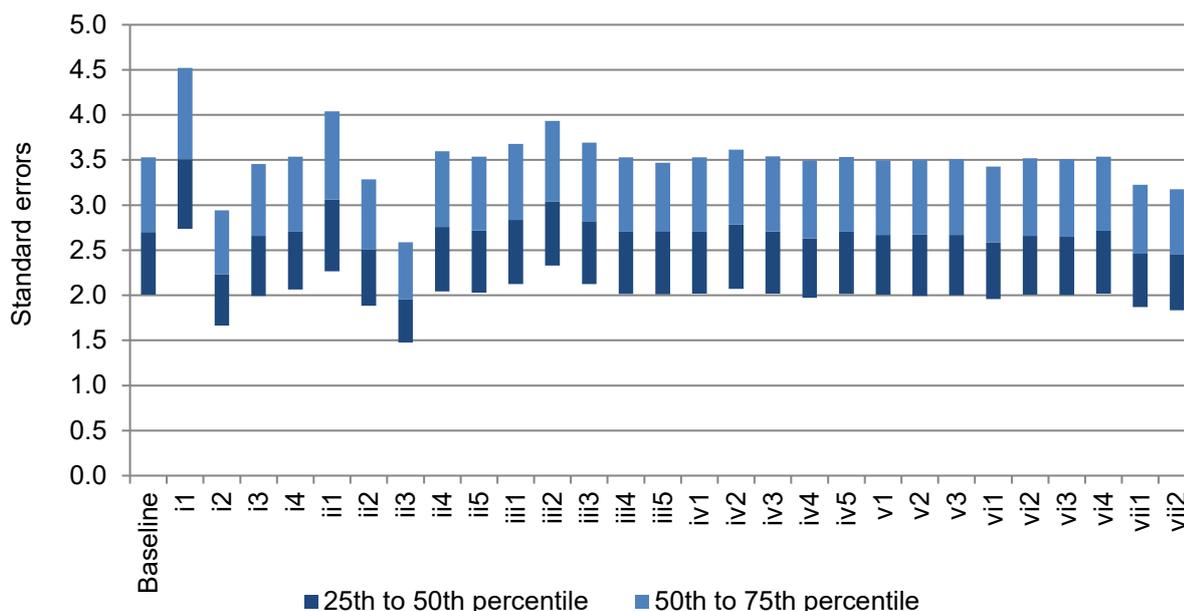
Note: Includes all courses for which we report estimates. The size of each bubble represents the number of FPE students on each course. Business and management courses are highlighted in red, engineering courses in blue, nursing courses in green and performing arts courses in orange.

Table 15: Decomposition of variance in actual earnings / indicator value

	Baseline estimates	B3 progression	Proportion earning <£3,000	Proportion earning below NLW
Variance of benchmark	71%	69%	48%	63%
Variance of difference from benchmark	16%	18%	54%	35%
Covariance between benchmark and difference	13%	13%	-2%	2%

Note: Covers all full-time first degree courses for which we report estimates. Weighted by full-person equivalents.

Figure 37: Distribution of standard errors on estimates, for baseline and alternative specifications



Note: Specifications are grouped as in Section 8 and numbered in the order they appear in each table in that section. Bars show the interquartile range of standard errors on estimates for courses we would report for each specification and are weighted by the number of students by course in the baseline specification.

Table 16: Impact of different criteria and thresholds on courses flagged for regulatory action

Difference from benchmark	No. of flagged courses	Proportion of graduates from flagged courses among:				Average selectivity percentile of flagged courses
		All graduates	FSM-eligible graduates	Graduates at London providers	Graduates of courses below B3 progression threshold	
Below –5ppt	499	23.9%	30.6%	20.9%	53.2%	35 th
Below –10ppt	241	9.4%	13.5%	8.6%	26.6%	29 th
Below –20ppt	28	0.8%	1.2%	0.4%	0.5%	23 rd
Below –10ppt & average earnings below £27,000	182	7.2%	10.7%	6.3%	26.6%	29 th
Significantly negative	380	21.8%	27.8%	15.6%	48.4%	35 th
Significantly below –5ppt	148	7.4%	10.5%	5.2%	25.3%	29 th
Significantly below –10ppt	44	1.7%	2.2%	0.4%	2.6%	26 th

Note: Full-time first degree courses only. Proportion of all graduates is of all FPE graduates from that level and mode, including those who do not meet the sample restrictions, or who are on courses for which we do not report results. ‘London providers’ are those with a registered address in the London or Slough and Heathrow TTWAs. Proportion of students on courses significantly below numerical B3 progression threshold includes only those where estimates are reported under our baseline and are available for the same subject area for 2017–20 from OfS data. Selectivity percentile reflects the rank of a course’s selectivity (measured by average GCSE points of graduates) within a bucket. Last three rows use a 95% confidence interval.

Table 17: Impact of different methodological variations on courses for which difference from benchmark is below –10 percentage points

Difference from benchmark	No. of flagged courses	Proportion of graduates from flagged courses among:				Average selectivity percentile of flagged courses
		All graduates	FSM-eligible graduates	Graduates at London providers	Graduates of courses below B3 progression threshold	
Baseline approach	241	9.4%	13.5%	8.6%	26.6%	29 th
Adjust earnings for local living costs (v3)	189	7.8%	12.0%	14.1%	27.7%	31 st
Hybrid random / fixed effects approach, with controls for selectivity (vi2)	114	4.1%	3.2%	2.4%	4.9%	59 th
Plus control for whether HEP in London (vi3)	108	3.8%	3.3%	4.7%	4.9%	58 th

Note: See note to Table 16. Specifications are as described in Section 8.

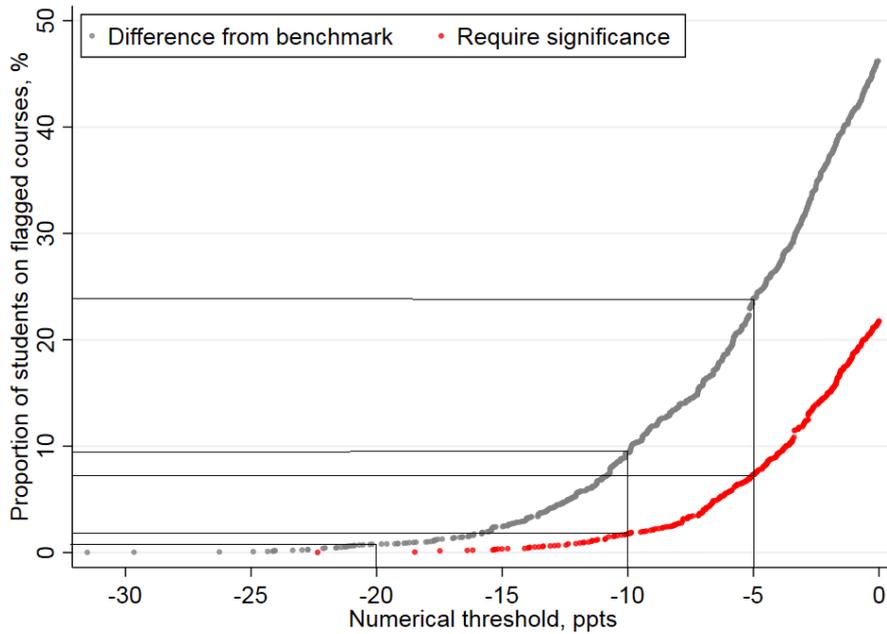
Table 18: Impact of different criteria and thresholds on courses flagged for regulatory action, by broad subject area

	Agri- culture	Allied to medicine	Archi- tecture	Bio-sci- ences & sports sci.	Business	Combined	Comms	Computing	Creative & perf. arts	Education	Engineer- ing and tech.
[1] All courses	42	209	53	169	142	14	90	101	260	91	116
[2] Courses for which we report estimates	19	141	41	129	96	3	59	74	183	60	58
	81%	93%	93%	96%	95%	37%	92%	93%	95%	94%	83%
Below –5ppt	11	29	14	31	42	2	14	28	50	15	16
	40%	12%	24%	20%	39%	18%	21%	40%	21%	10%	24%
Below –10ppt	8	17	12	8	21	0	4	16	14	8	9
	30%	5%	19%	3%	14%	0%	3%	19%	6%	4%	15%
Below –20ppt	0	4	3	0	3	0	0	2	1	0	2
	0%	1%	4%	0%	1%	0%	0%	4%	0%	0%	2%
Below –10ppt and average earnings below £27,000	8	10	11	8	18	0	4	5	14	8	5
	30%	3%	17%	3%	12%	0%	3%	7%	6%	4%	7%
Significantly nega- tive	8	23	12	16	43	0	9	21	37	14	13
	30%	11%	21%	13%	44%	0%	18%	36%	21%	12%	22%
Significantly below –5ppt	2	13	10	5	21	0	0	7	11	4	7
	11%	4%	17%	3%	17%	0%	0%	9%	7%	3%	13%
Significantly below –10ppt	1	4	3	0	7	0	0	2	2	1	3
	6%	1%	4%	0%	4%	0%	0%	4%	0%	1%	4%

	Geography	History & philosophy	Language & area studies	Law	Maths	Medicine	Physical sciences	Psychology	Social sciences	Veterinary
[1] All courses	61	144	150	91	62	27	108	102	313	11
[2] Courses for which we report estimates	50	96	121	78	50	26	58	86	199	5
	96%	91%	95%	96%	95%	100%	72%	97%	87%	46%
Below –5ppt	24	38	32	39	19	1	14	21	57	2
	36%	23%	20%	43%	28%	6%	19%	22%	23%	9%
Below –10ppt	8	18	14	25	12	0	7	5	33	2
	6%	10%	7%	27%	17%	0%	8%	4%	11%	9%
Below –20ppt	2	0	1	1	0	0	0	0	7	2
	2%	0%	0%	1%	0%	0%	0%	0%	2%	9%
Below –10ppt and average earnings below £27,000	8	18	14	23	0	0	5	5	16	2
	6%	10%	7%	24%	0%	0%	6%	4%	6%	9%
Significantly negative	19	22	24	35	12	1	6	16	47	2
	32%	17%	17%	43%	17%	6%	9%	18%	19%	9%
Significantly below –5ppt	5	7	4	19	9	0	1	3	18	2
	4%	6%	3%	22%	13%	0%	2%	2%	7%	9%
Significantly below –10ppt	2	1	2	4	1	0	0	0	9	2
	2%	1%	1%	3%	1%	0%	0%	0%	3%	9%

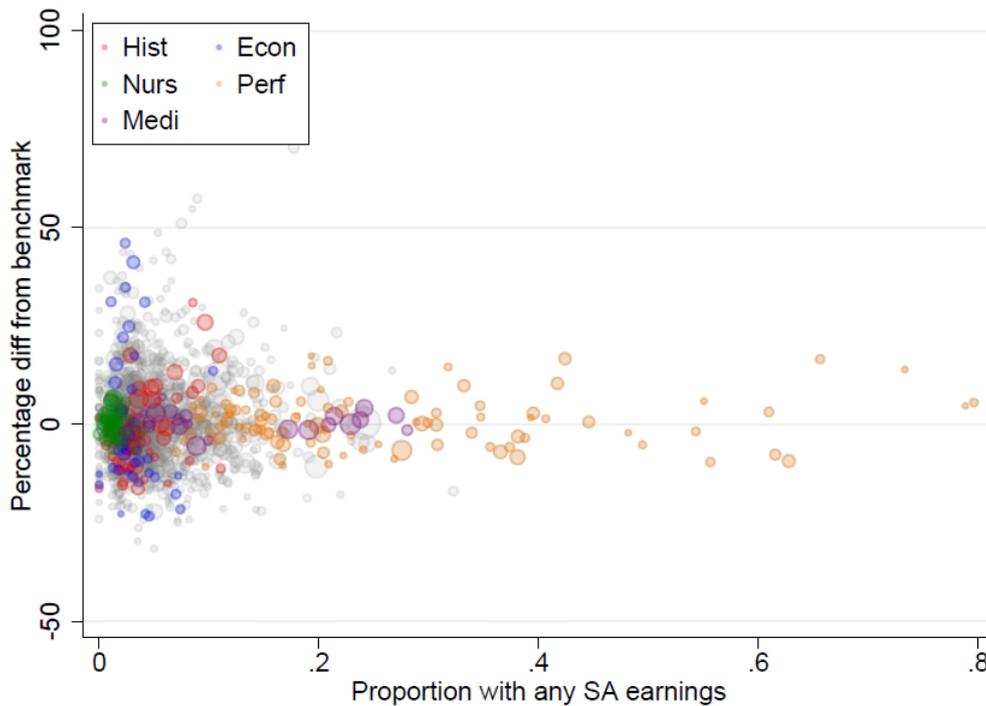
Note: Full-time first degree courses only. [1] includes all courses that have at least 10 FPE graduates in the sample. [2] includes those that meet our sample size restriction of at least 50 FPE graduates. Numbers are numbers of courses affected, and percentages are proportions of all FPE graduates who are on those courses among all graduates in a given subject area. Courses are grouped by the broadest subject area classification ('CAH level 1'). See <https://www.hesa.ac.uk/collection/coding-manual-tools/hecoscahdata/cah> for details. Last six rows use a 95% confidence interval.

Figure 38: Proportion of students on courses that would be flagged as being below different numerical thresholds under our baseline approach



Note: Full-time first degree courses only. 'Difference from benchmark' shows the result if the threshold was simply applied to the point estimates and 'Require significance' shows what would happen if the whole 95% confidence interval was required to be below the threshold.

Figure 39: Difference from benchmark and proportion of students with any self-assessment (SA) earnings in their highest-earning year three–five years after graduation, by course



Note: Includes all courses for which we report estimates. The size of each bubble represents the number of FPE students on each course.

References

- Arteaga, C. (2018). The effect of human capital on earnings: evidence from a reform at Colombia's top university. *Journal of Public Economics*, 157, 212–225.
- Belfield, C., Britton, J., Buscha, F., Dearden, L., Dickson, M., van der Erve, L., Sibieta, L., Vignoles, A., Walker, I., & Zhu, Y. (2018). The impact of undergraduate degrees on early-career earnings. Department for Education, Research Report DFE-RR808.
- Britton, J., Dearden, L., van der Erve, L., & Waltmann, B. (2020). The impact of undergraduate degrees on lifetime earnings. Institute for Fiscal Studies, Report R167.
- Britton, J., van der Erve, L., Belfield, C., Vignoles, A., Dickson, M., Zhu, Y., Walker, I., Dearden, L., Sibieta, L., & Buscha, F. (2022). How much does degree choice matter? *Labour Economics*, 79, 102268.
- Britton, J., van der Erve, L., Waltmann, B., & Xu, X. (2021). London calling? Higher education, geographical mobility and early-career earnings. Institute for Fiscal Studies, Report
- Dale, S. B., & Krueger, A. B. (2002). Estimating the payoff to attending a more selective college: an application of selection on observables and unobservables. *Quarterly Journal of Economics*, 117(4), 1491–1527.
- Department for Education (2017). Teaching Excellence and Student Outcomes Framework specification: October 2017.
- Mountjoy, J., & Hickman, B. R. (2021). The returns to college(s): relative value-added and match effects in higher education. National Bureau of Economic Research, Working Paper 29276.
- Office for Students (2018a). TEF subject pilot metrics: specification and rebuild document. January 2018.
- Office for Students (2018b). Teaching Excellence and Student Outcomes Framework: Year Four procedural guidance.
- Oster, E. (2019). Unobservable selection and coefficient stability: theory and evidence. *Journal of Business & Economic Statistics*, 37(2), 187–204.



Department
for Education



© IFS 2025

ISBN: 978-1-83870-647-0

The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

For any enquiries regarding this publication, contact us at:
jamie.burton@education.gov.uk or www.education.gov.uk/contactus