



Governments harnessing the power of data to get 'value for money': a simulation study of England's Office for Students B3 Proceed Metric

Alex Bradley & Martyn Quigley

To cite this article: Alex Bradley & Martyn Quigley (2023): Governments harnessing the power of data to get 'value for money': a simulation study of England's Office for Students B3 Proceed Metric, Studies in Higher Education, DOI: [10.1080/03075079.2023.2196292](https://doi.org/10.1080/03075079.2023.2196292)

To link to this article: <https://doi.org/10.1080/03075079.2023.2196292>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 04 Apr 2023.



Submit your article to this journal [↗](#)



Article views: 136





View related articles [↗](#)



View Crossmark data [↗](#)

Governments harnessing the power of data to get ‘value for money’: a simulation study of England’s Office for Students B3 Proceed Metric

Alex Bradley ^a and Martyn Quigley ^b

^aSchool of Education and Sociology, University of Portsmouth, Portsmouth, UK; ^bSchool of Psychology, Swansea University, UK

ABSTRACT

The mass participation in higher education has led to greater spending by governments and students which has increased the focus on graduate outcomes. In England, the Office for Students (OfS) is planning to take regulatory action, using the Proceed metric, against universities and their courses which do not have 60% of students with positive outcomes within 15 months of graduation. This study uses simulations to explore how effectively the Proceed metric can (a) identify the true population level of positive outcomes, (b) explore the precision of those estimates, and (c) accurately classify courses below the 60% threshold. The simulation varied: level of positive outcomes within the population (20–90%), sample size (40–1000), and the percentage of the population sampled (30–90%). The bias (difference between sample and population estimate), coverage probability (proportion of true population estimates within the intervals), and precision of confidence intervals (average range of confidence intervals) were calculated. Three main findings were (a) the Proceed metric is accurate in terms of bias and coverage probability, (b) the estimates lack precision, especially with small sample sizes, and (c) the imprecision will impact the Proceed metric ability to correctly classify courses below the 60% threshold. Governments seeking to collect graduate survey data and use it to regulate universities should make every effort to maximise sample size and should resist the temptation to regulate at a micro-level (i.e. courses) since classification will likely be inaccurate, especially for those just below the threshold.

ARTICLE HISTORY

Received 8 July 2022
Accepted 22 March 2023

KEYWORDS

Employability; OfS; B3 regulations; proceed metric; graduate outcomes

Introduction

Many countries around the world have seen increased participation in Higher Education (HE) which has resulted in larger debts for both governments and students (Bondar et al. 2020; OECD 2017). The increased spending on HE has led to a focus on the value for money of the university experience with a particular focus on graduate outcomes (Nghia et al. 2020). Governments around the world are increasingly looking to measure and regulate universities to ensure a return on their investments (European Commission/EACEA/Eurydice 2018). One example of such regulation is England’s Office for Students (OfS) Proceed metric which requires all students on a course to have over 60% of graduates in a graduate-level job within 15 months otherwise regulatory action could include a monetary

CONTACT Alex Bradley  alexander.bradley@port.ac.uk  School of Education and Sociology, University of Portsmouth, Portsmouth, PO1 2HY, UK

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

penalty, revoking a providers degree awarding powers or even deregistering a provider (Office for Students (OfS) 2022). This research aims to explore the bias in estimating the number of graduates in graduate-level jobs and how accurately it can determine whether a course is above the 60% threshold. Whilst the findings of this study will be particularly relevant to England, it will also provide valuable insights to other governments who may wish to follow a similar trajectory in using graduate survey data to regulate universities. We shall now review the rationale for mass participation in HE, provide an overview of international approaches to measuring and evaluating employability, delineate in more detail England's approach to measuring and regulating employability, and outline the current study.

The rationale for mass participation in HE

From a government perspective, HE is seen as a way of increasing the knowledge and skills of its workforce to be able to drive innovation, and productivity (Australian Government Department of Education 2020; Business Innovation Department 2015). For example, Holland et al. (2013) find that a 1% increase in the proportion of the workforce with a degree leads to a .2–.5% increase in long-run productivity. There are negative consequences to mass participation in HE with the obvious one being the cost of financing HE. For example, the average Organisation for Economic Co-operation and Development (OECD) country spends 1.4% percent of its gross domestic product per year on HE (OECD 2021). Another concern is over education where graduates with a degree become employed in a role that does not require a degree or over skilling where graduates get a role that does not require the skills developed in their degree program (Angeloni 2021). In both cases, governments and students have a less favourable return on the investments they have made in education (Dolton and Vignoles 2000; Quintano, Castellano, and D'Agostino 2008).

From a student perspective, HE is seen as attractive due to the graduate premium which is the idea that investing time and financial resources into attaining a university degree will lead to financial rewards in the future through accessing better-paid careers. For example, OECD (2017) finds that on average students with a bachelor's degree earn 48% more than those with school-level education. However, there is nontrivial variability in the lifetime earnings across different courses and different protected characteristics with some courses and groups receiving far less than others (Britton et al. 2020).

The exact nature of this relationship between HE and labour market outcomes is debated with two main competing theories: human capital theory and credentialism (Tomlinson 2008). Human capital theory argues that the process of completing HE endows individuals with capital which we can think of as knowledge, skills, expertise, etc. that improves that individuals capacity to be productive employees (Angeloni 2021). On the other, the credentialist perspective suggests that individuals compete with higher qualifications to stand out whilst businesses raise their entry requirements, so jobs that were previously done without a degree now require one (James et al. 2013). Both governments and students' have substantial costs from engaging in HE and therefore there is a focus on measuring and evaluating employability to ensure good outcomes.

International approaches to measuring and evaluating employability

Measuring employability

Measuring employability can be challenging with a variety of metrics to choose from each giving us different information about how well students transition from HE into the world of work. For example, across the European Higher Education Areas (EHEA) employability is monitored through unemployment rates, earnings of graduates, and the vertical mismatch of graduates (where the job does not require the qualification attained) (European Commission/EACEA/Eurydice 2018; Domínguez and Gutiérrez 2022). The way employability measures are captured falls into two broad categories administrative data or graduate self-report surveys. Several countries use graduate surveys.

For example, within the EHEA 38 countries have regular (i.e. France, Germany, and Austria) or ad-hoc nationally representative graduate surveys (i.e. Romania, Croatia, and Slovakia). However, graduate surveys have limitations, for example, they can be expensive, can have low response rates, sampling can be biased (i.e. only those who are unemployed respond) or small samples could lead to poor representation of the true population value (Schomburg 2011). Many of these limitations can be avoided by using administrative data. For example, within the UK, the Longitudinal Education Outcomes dataset provides insights into the earnings of graduates at one, three, and five years after graduation (Department for Education 2017). However, administrative data will never be able to provide insights into valuable questions like how satisfied graduates are with their careers. Policy-makers are increasingly keen to develop a graduate survey to allow them to more efficiently allocate funding to courses and universities with better graduate outcomes. One ongoing example of the increasing prominence of graduate outcomes survey is the European Graduate Tracking survey which aims to have 80% of member countries tracking their graduates by 2025 (European Commission 2021). The question we now turn to is how countries are using this data in their funding and regulation of universities.

Evaluating employability

Governments can incentivise universities and prospective students to improve their employability outcomes through a variety of different mechanisms which range from light information-based approaches (i.e. providing average salaries by course to prospective students) to stronger financial incentives/punishments; we review three strong mechanisms. The first approach utilizes employability measures to alter the funding of universities either by directly impacting the funding formula or via performance-based funding measures. For example, in Finland, the number of employed graduates impacts the funding received by universities and polytechnics in conjunction with other educational factors like the number of bachelor's degrees awarded, student credit, etc. (Boer et al. 2015). The second approach sees governments adjusting the number of places on courses or increasing the cost of courses depending on whether those skills are needed in the labour market. For example, Australia's job-ready graduate package aligns funding to university courses that are deemed key to areas of study that will contribute to national priorities (Australian Government Department of Education 2020). The third approach is to set targets for specific employability metrics like unemployment rates, number of graduate-level jobs, or earning levels of graduates and then take regulatory action against universities that do not meet those standards via the imposition of fines, removing student funding for courses, etc. However, if targets are based on self-report graduate surveys which have small samples and only reflect a small percentage of the target population, the estimate of true employability values could be inaccurate leading to universities facing financial and reputation damage unjustly.

England's context and approach to measuring and regulating employability

There are 170 higher education institutions in the UK with 2.66 million students in 2020–2021, the majority of which are undergraduates (1.94 million) (Universities UK 2022). Funding of HE is primarily through tuitions fees and to a less extent direct teaching grants from government. Most UK students can take out an income contingent tuition fee loan to pay for their studies and a maintenance grant to help with living costs which are repaid if they earn over a certain threshold in the future (Institute for Fiscal Studies 2022). Regulation of education and higher education is devolved in the UK with England, Scotland, Wales, and Northern Ireland having separate regulations and regulators. The focus of this work will be on the proposal by the Office for Students (OfS) who regulate English HE providers. The OfS has recently introduced several key reforms including B3 conditions of registration, a modified Teaching Excellence Framework (TEF) aimed at evaluating teaching with HE and Access and Participation Plans (APP) focused on widening participation with HE.

The primary focus of this research is OfS B3 conditions of registration for English universities that include the Proceed metric that requires a certain percentage of graduates within 15 months to be engaged in a positive outcome which primarily will mean graduate-level employment, or further study (Office for Students 2022a). The Proceed metric itself was first conceived back in 2020 by the OfS and has been modified following feedback from the sector and round table discussion with stakeholders to reach its present version (Office for Students (OfS) 2021). The Proceed's metric percentage with a positive outcome depends upon the mode of study (full-time, part-time, and apprenticeships) and levels of study (first degree, other undergraduates, etc.). So, for example, a full-time first-degree provision will need to attain 60% of graduates with a positive outcome whereas, full-time taught masters are expected to have 70% with a positive outcome (Office for Students 2022b). In addition, to the thresholds by mode and level of study, there are additional split indicators that they may choose to investigate like course, disability, ethnicity, sex, etc. For the current study, we focus on the 60% threshold for a course being studied full-time at the undergraduate level on their first degree since this represents a large proportion of the student population (i.e. 67% of the total student population in 20/21 (Higher Education Statistics Authority (HESA) 2022a)). The OfS will begin investigating universities when the 90% confidence interval is below the set threshold and will take regulatory action when the 95% confidence interval falls below and does not include the threshold (Office for Students 2022a, 62). Failure to hit the threshold with the 95% confidence interval could lead to the OfS (a) penalising the institution with a £500,000.00 fine, (b) restricting students' access to finance on courses, or (c) restricting universities' degree awarding powers and other potentially unpalatable actions for courses they regard as 'low quality' (Office for Students 2022a, 66). The implications for English universities which fall below the Proceed metric are clear and with more countries collecting graduate outcomes data (European Commission 2021) there will be a temptation for other policymakers to follow suit in regulating graduate outcome data. Therefore, the HE sector, policymakers, and politicians must understand the challenges of sample size, sampling, and misclassification that will undermine the effectiveness and legitimacy of such regulation.

Current study

The objective of this simulation is to explore the accuracy of the OfS Proceed metric in identifying students with positive outcomes and to determine how effectively it can identify courses below the 60% threshold. Simulation from a statistical perspective is a study that repeatedly generates synthetic data for a virtual population and then takes repeated random samples from that population under pre-specified conditions that can be varied across different simulation scenarios (Boulesteix et al. 2020). For example, we can generate a synthetic population of graduates of a certain size with a certain percentage of graduate-level employment. The strength of conducting a simulation study is it can provide insights into how biased the estimates of positive outcomes are (i.e. sample estimate vs true population percentage in graduate jobs), the coverage probability of confidence intervals (i.e. the probability that confidence interval will contain the true value), the precision of the confidence intervals (i.e. how wide are they), and the likelihood of correctly or incorrectly classifying courses below the 60% threshold (Morris, White, and Crowther 2019; Boulesteix et al. 2020). This research explores three crucial questions: first, how do factors like percentage sampled and sample size impact the accuracy of estimating the true level with positive outcomes? Second, how precise are the OfS confidence intervals? Third, what conditions of sample sizes and percentage sampled are required to correctly identify courses below the 60% threshold?

Method

Simulation description

This study is a simulation study designed to replicate the OfS B3 Proceed metric which is itself based on the data collected from the graduate outcomes survey. In our simulation, we vary three key

variables. First, the percentage of graduates in the population with a positive outcome varied from 20% up to 95% increase by 5% increments (i.e. 20%, 25%, ... , 95%). These levels were chosen as limits because anything less than 20% was not likely to be identified as above the 60% target and anything greater than 90% would be unlikely to be identified as below the threshold. Second, we vary the population size from 40 students up to 1000 students going up in increments of 10. These limits were chosen as it was thought to be unlikely that courses below 40 students would be included as the OfS does not use samples less than 23 students (Office for Students 2022c). Third, we varied the percentage sampled from the population from a minimum of 30%, as set by the OfS, to 90% going up in increments of 5% (i.e. 30%, ... , 90%)(Office for Students 2022d).

Simulation procedure

To create the simulation, we followed the following three steps: data generation, sampling of the data, and calculating statistics from each of the samples.

First data generation involved the creation of 97 data frames each with 16 variables. The 97 datasets went from 40 rows (i.e. student outcomes) in length up to 1000 rows long in ten increments per dataset (i.e. the second dataset had 50 rows). Each row symbolises a graduate with a positive outcome (1) or negative outcome (0). Each variable within a data frame represented a population with a certain percentage of positive outcomes. The first variable in a dataset had 20% of the rows with a 1 symbolising a population with 20% positive outcomes whilst, the second variable had 25% with a positive outcome and this percentage increased by 5% in each variable till the 16th variable which had 95% with a positive outcome.

The second step was to take a randomly selected sample, without replacement, from each of the 16 variables in each of the 97 datasets. This sample varied from 30% of the population to 90% of the population. This process was then repeated 100 times for each of the 16 variables in all 97 datasets.

Finally, from each of the samples, three statistics were calculated. First is the percentage with a positive outcome/graduate job. Second, the Jeffries 95% confidence interval and third, the Jeffries 90% Confidence interval. Jeffries confidence intervals were chosen by the OfS as they are known to have favourable properties when estimating intervals on binomial proportions and when used on small samples (Brown, Cai, and Dasgupta 2001; Office for Students 2022d). Due to OfS not using a student population of less than 23, any samples with less than this number were removed before analysis (20,800 cases were removed). The final dataset had 1,996,800 observations.

Simulations were performed using four packages ('tidyverse', 'DescTools', 'kableExtra','Metrics') with R Studio (version 1.4.1717) running R version 4.1.1. All the code and data are accessible at the Open Science Framework (Reviewer link: https://osf.io/rgxap/?view_only=737009e364f74351b7bace98e6461b92).

Simulation outcome measure

The performance of the Proceed metric was evaluated using the following four measures: bias, error in the coverage probability, precision, and misclassification.

Bias represents the extent to which the sample estimated percentage with a positive outcome varies from the population level with a positive outcome (see Figure 1) (Morris, White, and Crowther 2019). Coverage probability refers to the proportion of true population values captured within the confidence interval (Agresti and Coull 1998). Both bias and coverage probability give us a sense of the accuracy with which the OfS Proceed metric can correctly identify the actual level of graduates with a positive outcome. Precision refers to the average range of the confidence intervals (Montori et al. 2004). Misclassification in this instance will refer to either when the confidence intervals for a population with less than 60% with a positive outcome crossed the 60% threshold or when the confidence interval for a population with 60% (i.e. false positive) or greater positive outcomes and confidence intervals that are below that threshold (i.e. false negative). For example, Figure 1 has a

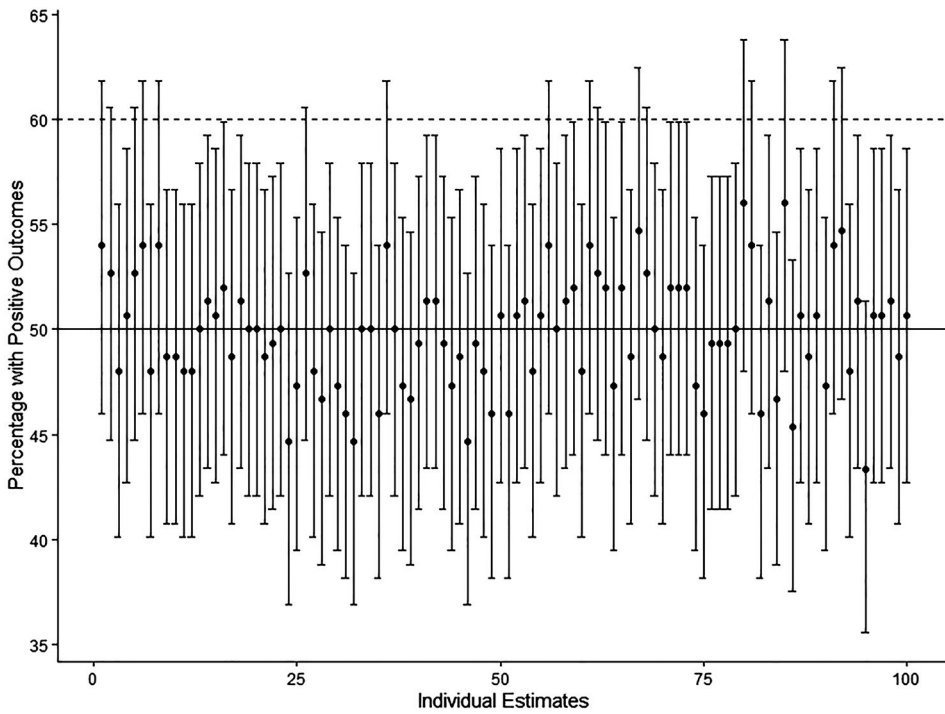


Figure 1. This graph depicts 100 estimates from a population with 50% positive outcomes and a sample size of 150 students. Bias is captured by the distance from each point to the bold line. Precision is the average range of the confidence intervals (distance between the top and the bottom of the confidence interval).

population of 50% with positive outcomes and shows no sign of misclassification since all confidence intervals fall below the 60% threshold. Finally, all standard deviations reported refer to 1 standard deviation above or below.

Results

Bias and coverage probability by percentage sampled and sample size

Table 1 shows that the mean level of bias is very small indicating that the sample estimate is very close to the true population value across all levels of the percentage sampled. Coverage probability

Table 1. Shows that the average level of Bias (difference between sample and population estimate) was generally small and became smaller as percentage sample increased. The Coverage Probability remained close the expected 95% and 90% confidence intervals across the percentage sampled.

% Sampled	Number of data points in Simulation	Bias	SD	Coverage Probability 95% CI	Coverage Probability 90% CI
30	148800	-0.06	4.23	94.84	89.92
35	150400	-0.05	4.02	94.86	89.79
40	152000	-0.06	3.85	94.91	89.92
45	152000	-0.05	3.62	94.98	89.96
50	153600	-0.04	3.52	95.15	90.06
55	153600	-0.04	3.38	95.09	89.93
60	155200	-0.01	3.30	95.17	90.11
65	155200	-0.02	3.19	95.12	90.02
70	155200	-0.01	3.07	95.03	89.93
75	155200	-0.01	2.96	95.11	90.07
80	155200	-0.02	2.86	95.13	90.14
85	155200	-0.02	2.79	95.11	90.12
90	155200	-0.01	2.70	95.15	90.20

was good at around 95% for the 95% confidence intervals and fluctuating around 90% for the 90% confidence intervals as would be expected. This did not seem to change much with varying levels of the percentage sampled.

Table 2 illustrates bias is within a percent of the true population level and generally becomes less biased as the sample size increases up to around 350. Equally, the standard deviation of bias reduces as sample size increases with less variability within samples over 100. The coverage probability appears to fluctuate randomly around the 95% mark with the 95% confidence intervals and around 90% for the 90% confidence as would be expected. So, large sample sizes reduce bias and especially the variability in the bias estimates but do not appear to cause an issue for coverage probability.

The precision of confidence intervals by different levels of percentage sampled and sample size

Table 3 shows that the average range of 95% confidence intervals goes from 14.96% to 9.35% reducing as the percentage sample size increases. Whilst the average range of 90% confidence intervals varies from 12.59% to 7.86%. This large average range for the confidence interval, especially the 95%, will likely impact the OfS's ability to discern between courses above or below the 60% threshold which we explore in more detail in the next section.

Table 4 highlights that samples of less than 50 students will produce imprecise confidence intervals with an average range of 26% for 95% confidence intervals and 22.59% wide for 90% confidence intervals. When samples increase over 100 the average confidence interval has reduced to 14.93% for 95% confidence intervals and 12.55% for 90% confidence intervals. When samples increase to above 300, the width of the 95% confidence interval falls below 10% for both the 95% and 90% confidence intervals. The relation between sample size and confidence intervals is highlighted in Figure 2. So, whilst variability in confidence intervals does reduce as the percentage sampled increases the sample size appears to be a key factor with samples below 100 having particularly large confidences which will reduce the OfS's ability to accurately discern whether courses have met the 60% criteria.

Correctly classifying population scores of graduates with less than 60% attaining a positive outcome using 95% and 90% confidence intervals

Table 5 shows with a population level of 45% attaining a positive outcome 5.95% of cases will have an upper 95% confidence interval above the 60% threshold. When at a population level, there are

Table 2. Illustrates that bias (difference between sample and population estimate) and the error around the bias reduces as sample sizes increase. Coverage probability remains around the expected levels for the 95% and 90% confidence intervals.

Sample size	Number of data points in Simulation	Bias	SD	Coverage Probability 95%	Coverage Probability 90%
24–50	102400	0.79	7.32	95.14	89.89
51–100	195200	0.06	5.15	94.96	89.90
101–150	193600	–0.21	3.91	94.98	90.02
151–200	193600	–0.18	3.29	95.06	90.04
201–250	192000	–0.18	2.93	95.02	89.75
251–300	196800	–0.13	2.66	94.85	89.68
301–350	164800	–0.09	2.43	95.00	89.99
351–400	147200	–0.05	2.24	95.31	90.16
401–450	126400	0.03	2.09	95.51	90.38
451–500	102400	0.00	1.99	95.33	90.27
501–550	92800	0.03	1.90	94.91	90.02
551–600	78400	0.03	1.82	94.90	90.26
601–650	62400	–0.01	1.78	94.49	89.76
651–700	51200	–0.04	1.71	94.85	89.76
701–750	40000	–0.04	1.63	95.22	90.41
751–800	28800	–0.01	1.57	95.46	90.92
801–850	19200	–0.01	1.51	95.54	91.25
851–900	9600	–0.04	1.53	94.82	89.80

Table 3. Demonstrated how as Percentage sampled increase the precision of the confidence intervals improves with smaller ranges with less variability in the range which is true for both the 95% and 90% confidence interval.

% Sampled	Number of data points in Simulation	Average range of 95% CI	SD of range of 95% CI	Average range of 90 CI	SD of range of 90% CI
30	148800	14.96	6.07	12.59	5.14
35	150400	14.10	5.98	11.86	5.07
40	152000	13.38	5.91	11.26	5.00
45	152000	12.66	5.62	10.64	4.75
50	153600	12.23	5.74	10.28	4.86
55	153600	11.69	5.52	9.83	4.66
60	155200	11.40	5.70	9.59	4.82
65	155200	10.98	5.50	9.23	4.65
70	155200	10.58	5.30	8.89	4.48
75	155200	10.24	5.15	8.60	4.35
80	155200	9.91	4.98	8.33	4.21
85	155200	9.62	4.85	8.09	4.09
90	155200	9.35	4.71	7.86	3.98

50% with a positive outcome and then 20% of cases will have 95% upper confidence interval above the threshold. At a population level of 55% with a positive outcome then 57% of cases will have upper confidence above the 60% threshold, so over half of these cases will not meet the criteria of strong regulatory action and 45% will not be investigated at the 90% confidence interval. On the other side, there will be 3.41% of cases where the population level with a positive outcome is at the 60% threshold yet their upper 95% confidence interval will suggest they are below the 60% threshold. So, University 'X' could find themselves below the 60% threshold due to sampling when in fact, the unknown population level of positive outcomes was 60% whereas, University 'Y' could have 55% with positive outcomes yet due to sampling error the confidence intervals do not correctly identify it as below.

Correct classification of courses below 60% by percentage sampled and sample size

Figure 3 highlights that as the percentage of graduates with a positive outcome in the population gets closer to the threshold the larger the percentage sampled must be to correctly identify those below the threshold. For example, when the level of graduates with positive outcomes in the

Table 4. Shows how smaller sample sizes can have quite imprecise confidence intervals with a large range which becomes more precise as sample size increases which is true for both the 95% and 90% confidence intervals.

Sample Size	Number of data points in Simulation	Average range of 95% CI	SD of range of 95% CI	Average range of 90 CI	SD of range of 90% CI
24–50	102400	26.76	6.20	22.59	5.32
51–100	195200	19.18	4.16	16.15	3.53
101–150	193600	14.93	2.94	12.55	2.48
151–200	193600	12.66	2.40	10.64	2.02
201–250	192000	11.19	2.09	9.40	1.76
251–300	196800	10.12	1.88	8.50	1.58
301–350	164800	9.31	1.73	7.82	1.45
351–400	147200	8.67	1.61	7.28	1.35
401–450	126400	8.14	1.50	6.83	1.26
451–500	102400	7.71	1.42	6.47	1.19
501–550	92800	7.34	1.35	6.16	1.13
551–600	78400	7.01	1.28	5.88	1.08
601–650	62400	6.72	1.23	5.64	1.04
651–700	51200	6.47	1.18	5.43	0.99
701–750	40000	6.24	1.14	5.24	0.96
751–800	28800	6.04	1.10	5.07	0.92
801–850	19200	5.86	1.06	4.92	0.89
851–900	9600	5.68	1.04	4.76	0.87

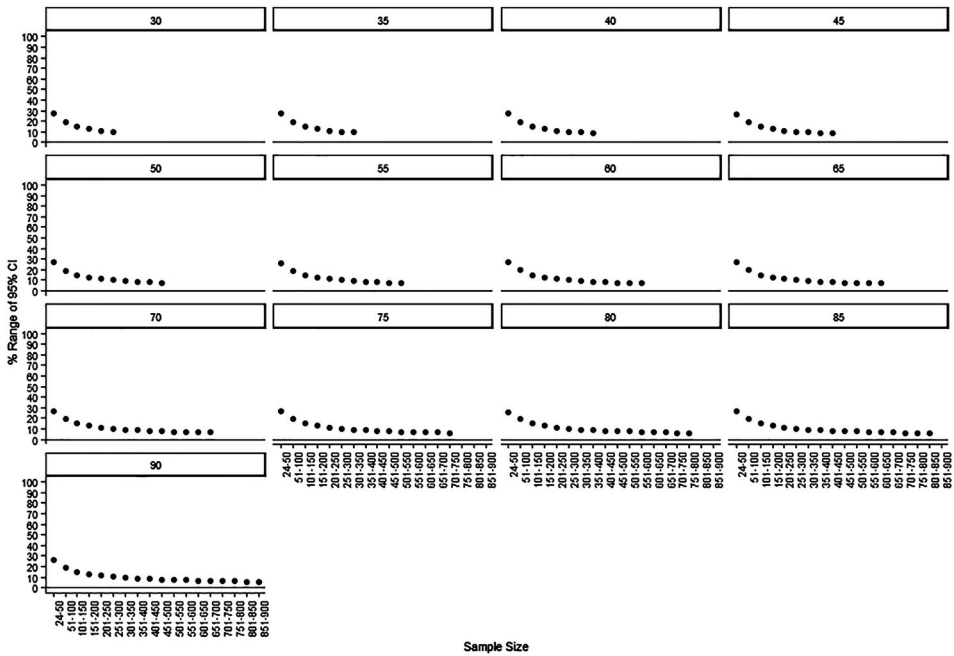


Figure 2. These graphs show reductions in the width of 95% confidence intervals as sample size increases at varying levels of the percentage sampled. Each box presents a certain percentage of the absolute population sampled from 30% to 90%.

population is at 40% or below then the percentage with upper confidence above 60% remains below 5%. When there are 50% within the population with a positive outcome then with 30% sampled on average 33% would not be identified as being below the threshold, and even when 90% of the population is sampled 13% would be classified as not having strong enough evidence of being below the threshold for regulatory action. With 55% of positive outcomes in the population then at 30% of the population sampled there would be 73% of cases with an upper confidence interval above 60%. This reduces to 43% not being identified as being below 60% when 90% of the population is sampled.

Figure 4 demonstrates that small sample sizes of 100 or less can lead to an inability to accurately identify courses below the 60% threshold even when the graduate outcomes are as low as 40 or 45%

Table 5. Demonstrates that as the percentage in the population with graduate jobs gets closer to the 60% threshold the less likely they are to be classified below the 60% threshold.

% Population in Graduate Job	Number of data points in Simulation	% UCI 95 Less 60	% UCI 90 Less 60
20	124800	0.00	0
25	124800	0.06	0.02
30	124800	0.39	0.2
35	124800	1.13	0.76
40	124800	1.27	
45	124800	5.95	4.09
50	124800	20.05	15.35
55	124800	57.08	45.64
60	124800	96.59	93.68
65	124800	99.98	99.94
70	124800	100.00	100
75	124800	100.00	100
80	124800	100.00	100
85	124800	100.00	100
90	124800	100.00	100
95	124800	100.00	100

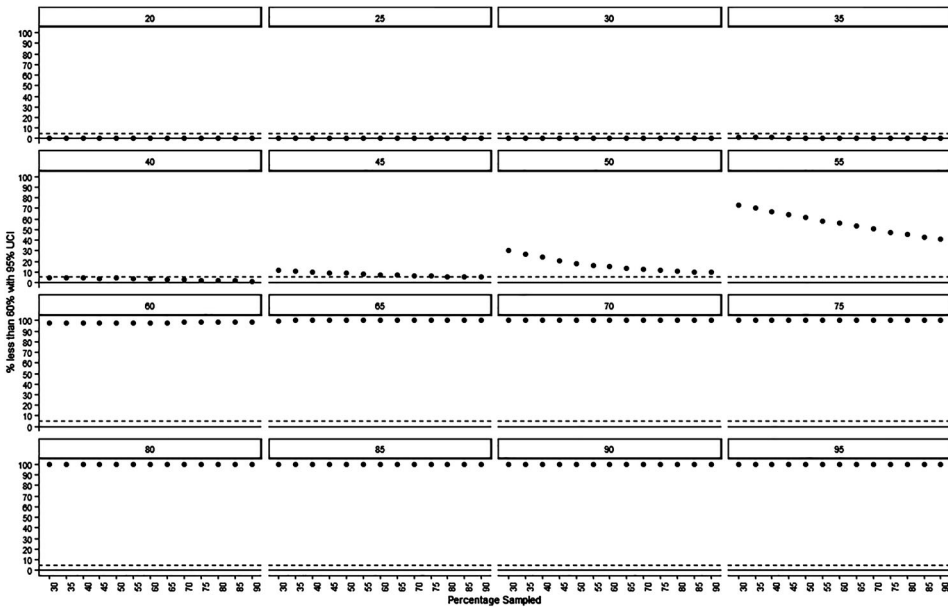


Figure 3. These graphs depict the percentage of cases where the upper 95% confidence interval will not be at or above the 60% threshold by the percentage of the population that is sampled. Each graph represents the percentage with a positive outcome/graduate job in the population starting from 20% top left to 95% bottom right. The dashed line represents the 5% level of error expected using 95% CI.

with positive outcomes. For example, when 45% have a positive outcome within the population with a sample of 50 or less there is a 56% chance it will not be correctly identified as being below the threshold. As the percentage with positive outcomes increases to 50% and 55% having a positive

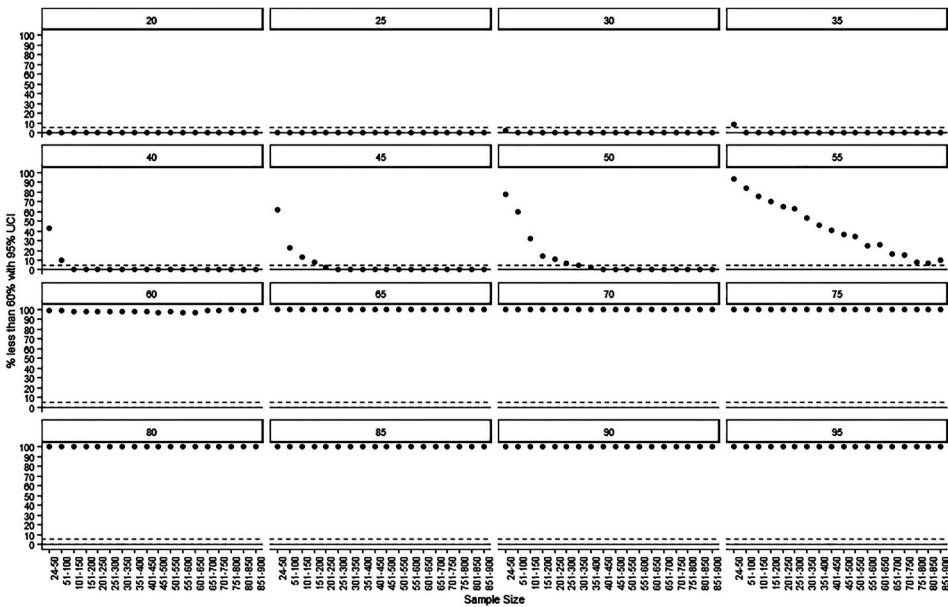


Figure 4. The graphs illustrate the percentage of cases where the upper 95% CI will not be at or above the 60% threshold by sample size that is sampled. Each panel indicates the percentage with a positive outcome in the population starting from 20% top left to 95% on the bottom right. The dashed line represents the 5% level of error expected using 95% CI.

outcome in the population then 84.77% and 94.5% of courses cannot be correctly identified at the 95% level as being below the threshold. The second point the graph makes is that as the percentage with positive outcomes gets closer to the 60% threshold increasingly large sample sizes are required to accurately identify those below the threshold. For example, with a sample of 51–100 then it is not until there are 45% with positive outcomes in the population that incorrect identification of courses starts to increase up to nearly a quarter of courses (24.66%) which then further increases to over half (57.88%) and over three quarters (87.43%) at 50% and 55% in graduate-level jobs. When sample sizes are over 100 then the majority (less than 5%) will be correctly classified as below 60% up until 50% of graduates have a positive outcome. At 50% of graduates with positive outcomes then a sample size of over 100 leads to 40% of courses not being correctly identified as below the 60% threshold. It is not until the sample reaches over 350 that the percentage of courses not identified as below the threshold drops to around 5%. When 55% of the population has a positive outcome then no sample size up to a 1000 drops below 10% of courses being correctly classified as below the threshold. This figure illustrates the challenges of correctly identifying courses that have fallen below the threshold (60%) when sample sizes are small and even if sample sizes are large once the population level reaches 50 or 55% with positive outcomes there will be substantial numbers of courses that cannot with confidence be identified as below the threshold.

Discussion

The current research explored how well the OfS Proceed metric could identify the true population level with a positive outcome and crucially, how accurate it is at distinguishing between those courses that fall below the 60% threshold and those courses that don't. These questions are of fundamental importance to English universities, the OfS, and the English government as there are substantial financial and reputational consequences should the Proceed metric be inaccurate or unable to correctly identify courses that fall below the target. Within this context, this study makes three important contributions to the efficacy of using graduate survey data to regulate universities. First, the Proceed metric does appear to be accurate at a population level, in terms of bias and coverage probability, even when the sample size and the percentage sampled are small. Second, the precision of these estimates is relatively imprecise especially when sample sizes and the percentage sampled are small. Third, this lack of precision means that it does not distinguish well between courses that are below the 60% threshold. We shall now review each of these three findings, highlight wider issues with the Proceed metric, and outline limitations within the present simulation.

First, the accuracy of the Proceed metric regardless of sample size and percentage sampled is good. The coverage probability fluctuated closely to the 95% and 90% levels for both the 95% and 90% confidence intervals. This illustrates that most of the time the true population of positive outcomes will be included within the lower and upper bounds of the confidence interval in most samples irrespective of the sample size and percentage sample. It also supports the idea that graduate surveys can be used by other countries to accurately identify the true population value of graduates with positive outcomes from relatively small samples. However, it ought to be noted that the Proceed metric only provides information primarily on whether students have achieved professional-level work it does not tell us anything about the quality of the work they do in terms of employment, work-life balance, job design (use of skills, progression opportunities, etc.), social support and other factors of good quality work (CarnegieUk Trust 2018). The Proceed metric also imposes a measure of what is successful based upon whether the job is deemed professional as opposed to taking a student-centred approach what they deem to be a success through whether they find the work meaningful and whether it fits in with their life goals.

Second, the confidence interval can on average be relatively wide meaning that the true value could fall between a large range of numbers. Whilst, increasing the percentage of the population sampled doesn't help to reduce the precision increasing the sample size does with sample sizes over 300 with an average range of confidence around 9%. Previous research has found that the

Jeffrey interval can be wider (i.e. less precise) than other intervals (Franco et al. 2019; Dean and Pagano 2015). Meaning that changing from the Jeffries confidence interval to another method of calculating confidence intervals like the Wilson's could increase the precision of confidence which would ultimately lead to more accurate identification of courses below the 60% threshold. Future simulation studies are required to explore the best types of confidence interval to use with this data.

Third, the imprecision within the confidence intervals makes it difficult to identify courses that are well below the 60% threshold. For example, when only 35% of the population has a positive outcome and the size of the course is 50 or less then 21% of the time it will not be identified as below the threshold. Furthermore, the closer that the true population value gets to the threshold the larger the sample size required to correctly identify courses below the target. For example, with 50% of the population with a positive outcome, a sample of over 300 is required to ensure that 90% of courses are identified as having not met the target.

Limitations

No simulation is definitive and whilst it can shed light on important issues like accuracy, precision, and misclassification it can't account for other factors that will influence estimates (Morris, White, and Crowther 2019). This simulation cannot speak to the important issue around what should and should not qualify as a graduate-level job. For example, a senior care home worker who provides support and comfort to those who are dying would be classified as a non-graduate level position whereas a sales accounts and business development manager would be a graduate-level position. One provides invaluable comfort to people and has obvious social benefits whilst the other looks at sales data and handles customer accounts.

Conclusions

This research has shown that a graduate outcome survey can when conducted appropriately provide an accurate estimate of the true population level with positive outcomes. However, when sample sizes are small and/or the percentage sampled is small these estimates lack precision which in turn impacts the regulator's ability to correctly classify courses as below the threshold. Several studies suggest one reliable way to increase response rates is through offering financial incentives (Church 1993; Singer et al. 1999) which ought to be considered given response rates to the Graduate Outcome survey have been 52% or below for the last three iterations of the survey (Higher Education Statistics Authority (HESA) 2022b). These findings have important implications for the OfS and English universities which are undergoing the first Proceed assessment this academic year. For instance, it would be unwise for universities to decide to close small courses where positive outcomes fluctuate around 50% or 55% since they are unlikely to be identified under the 60% threshold. From an international perspective, countries that plan on developing graduate outcome surveys should attempt to design and collect data to maximise sample size and percentage sampled and should avoid attempts at regulating at a micro-level especially if thresholds are to be applied. Even more radically future regulation could avoid the challenges highlighted by this simulation around sample size, precision, and misclassification if regulation focused less on student outcomes but on processes universities invested in enhancing their student employability. For example, employability initiatives within subject areas, university employability strategies, employer engagement within the curriculum. Alternatively, countries that do use graduate outcome measures would also do well to remember that the benefits of HE extend well beyond attaining professional-level jobs with HE also leading to citizens being more like to vote, lead healthier lifestyles, engage in more volunteering, and provide more educational activities for their children (Ma, Pender, and Welch 2016). A more comprehensive measure of the value of HE would be holistically designed to capture a broad range of outcomes beyond professional-level employment.

Acknowledgements

AB is the guarantor. AB created the simulator data and analysed the data. AB and MQ drafted the manuscript. All authors read, provided feedback, and approved the final version of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Data availability statement

Data and code to analyse data will be available from the Open Science Framework. The material is currently accessible for peer review using this link **HERE**.

Code availability

The dataset and code to reproduce the results will be available on the Open Science Framework and are available for reviewers **HERE**.

Ethics approval

No ethics approval was required since no data was collected.

Consent to publish

All authors approved the final manuscript before it was sent off for review.

ORCID

Alex Bradley  <http://orcid.org/0000-0003-4304-7653>

Martyn Quigley  <http://orcid.org/0000-0003-4342-1369>

References

- Agresti, Alan, and Brent A. Coull. 1998. "Approximate Is Better Than "Exact" for Interval Estimation of Binomial Proportions." *American Statistician* 52 (2): 119–26. doi:10.2307/2685469.
- Angeloni, Silvia. 2021. "A Policy Reform Aimed at Improving the Job Quality of Graduates." *Higher Education Policy* 34 (4): 861–80. doi:10.1057/s41307-019-00169-7
- Australian Government Department of Education, Skills and Employment. 2020. 'Job-Ready Graduates'.
- Boer, Harry de, Ben Jongbloed, Paul Benneworth, Leon Cremonini, Renze Kolster, Andrea Kottmann, Katharina Lemmens-Krug, and Hans Vossensteyn. 2015. 'Performance-Based Funding and Performance Agreements in Fourteen Higher Education Systems Report for the Ministry of Education, Culture and Science'. www.utwente.nl/cheps.
- Bondar, Tamara Ivanivna, Nataliia Viktorivna Telychko, Hanna Vasylivna Tovkanets, Tetiana Dmytrivna Shcherban, and Vasyl Ivanivych Kobal. 2020. "Trends in Higher Education in EU Countries and Non-EU Countries: Comparative Analysis." *Revista Romaneasca Pentru Educatie Multidimensionala* 12 (1Sup1): 77–92. doi:10.18662/rrem/12.1sup1/224
- Boulesteix, Anne Laure, Rolf H.H. Groenwold, Michal Abrahamowicz, Harald Binder, Matthias Briel, Roman Hornung, Tim P. Morris, Jörg Rahnenführer, and Willi Sauerbrei. 2020. "Introduction to Statistical Simulations in Health Research." *BMJ Open* 10 (12): 1–11. doi:10.1136/bmjopen-2020-039921.
- Britton, Jack, Lorraine Dearden, Laura van der Erve, and Ben Waltmann. 2020. 'The Impact of Undergraduate Degrees on Lifetime Earnings'.
- Brown, Lawrence, Tony Cai, and Anirban Dasgupta. 2001. "Interval Estimation for a Binomial Proportion." *Statistical Science* 16 (2): 101–33. doi:10.1214/ss/1009213285
- Business Innovation Department. 2015. 'UK Skills and Productivity in an International Context'.
- CarnegieUK Trust. 2018. 'Measuring Good Work'.
- Church, Allan H. 1993. "Estimating the Effect of Incentives on Mail Survey Response Rates: Meta-Analysis." *Public Opinion Quarterly* 57 (62): 62–79. <https://academic.oup.com/poq/article/57/1/62/1833464>. doi:10.1086/269355

- Dean, Natalie, and Marcello Pagano. 2015. "Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys." *Journal of Survey Statistics and Methodology* 3 (4): 484–503. doi:10.1093/jssam/smv024
- Department for Education. 2017. 'Employment and Earnings Outcomes of Higher Education Graduates by Subject and Institution: Experimental Statistics Using the Longitudinal Education Outcomes (LEO) Data'.
- Dolton, Peter, and Anna Vignoles. 2000. "The Incidence and Effects of Overeducation in the U.K. Graduate Labour Market." *Economics of Education Review* 19 (2): 179–198. doi:10.1016/S0272-7757(97)00036-8.
- Domínguez, Juan, and César Gutiérrez. 2022. "Bologna Process and Its Impact on Spanish Graduates Employability: Good News Yet to Come." *Higher Education Policy* May: 1–11. doi:10.1057/s41307-022-00274-0.
- European Commission. 2021. 'Towards a European Graduate Tracking Mechanism'. <https://doi.org/10.2766/67489>.
- European Commission/EACEA/Eurydice. 2018. 'The European Higher Education Area in 2018: Bologna Process Implementation Report'. Luxembourg. <https://doi.org/10.2797/63509>.
- Franco, Carolina, Roderick J.A. Little, Thomas A. Louis, and Eric V. Slud. 2019. "Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys." *Journal of Survey Statistics and Methodology* 7 (3): 334–64. doi:10.1093/jssam/smy019
- Higher Education Statistics Authority (HESA). 2022a. 'Who's Studying in HE?' 2022. <https://www.hesa.ac.uk/data-and-analysis/students/whos-in-he>.
- Higher Education Statistics Authority (HESA). 2022b. 'Graduate Outcomes 2019/20: Summary Statistics - Summary'. Summary Statistics Graduate Outcomes. 16 June 2022.
- Holland, Dawn, Iana Liadze, Cinzia Rienzo, and David Wilkinson. 2013. 'The Relationship between Graduates and Economic Growth across Countries'.
- Institute for Fiscal Studies. 2022. 'Higher Education'. 2022.
- James, Susan, Chris Warhurst, Gerbrand Tholen, and Johanna Commander. 2013. "What We Know and What We Need to Know About Graduate Skills." *Work, Employment and Society* 27 (6): 952–63. doi:10.1177/0950017013500116
- Ma, Jennifer, Matea Pender, and Meredith Welch. 2016. 'The Benefits of Higher Education for Individuals and Society About the Authors'.
- Montori, Victor M., Jennifer Kleinbart, Thomas B. Newman, Sheri Keitz, Peter C. Wyer, Virginia Moyer, and Gordon Guyatt. 2004. "Tips for Learners of Evidence-Based Medicine: 2. Measures of Precision (Confidence Intervals)." *Canadian Medical Association Journal* 611–615. doi:10.1503/cmaj.1031667.
- Morris, Tim P., Ian R. White, and Michael J. Crowther. 2019. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38 (11): 2074–2102. doi:10.1002/sim.8086
- Nghia, Tran Le Huu, Thanh Pham, Michael Tomlinson, Karen Medica, and Christopher D. Thompson. 2020. "The Way Ahead for the Employability Agenda in Higher Education." In *Developing and Utilizing Employability Capitals: Graduates' Strategies Across Labour Markets*, 256–76. Routledge. doi:10.4324/9781003004660-18.
- OECD. 2017. 'Benchmarking Higher Education System Performance: Conceptual Framework and Data, Enhancing Higher Education System Performance'.
- OECD. 2021. "What Is the Total Public Spending on Education?" In *Education at a Glance*, 268–281. Paris: OECD Publishing. doi:10.1787/0fab223e-en.
- Office for Students. 2022a. 'Consultation on a New Approach to Regulating Student Outcomes'.
- Office for Students. 2022b. 'Setting Numerical Thresholds for Condition B3'. <https://www.officeforstudents.org.uk/publications/setting-numerical-thresholds-for-condition-b3/>.
- Office for Students. 2022c. 'Supporting Information about Constructing Student Outcome and Experience Indicators for Use in OfS Regulation Description and Methodology'. <https://www.officeforstudents.org.uk/media/92b8b714-9a83-4817-b633-7c075ea17a40/description-and-methodology-document.pdf>.
- Office for Students. 2022d. 'Supporting Information about Constructing Student Outcome and Experience Indicators for Use in OfS Regulation: Description of Statistical Methods'. Supporting information about constructing student outcome and experience indicators for use in OfS regulation: Description of statistical methods.
- Office for Students (OfS). 2021. 'Projected Completion and Employment from Entrant Data (Proceed). Updated Methodology and Results.' www.officeforstudents.org.uk/publications/developing-an-
- Office for Students (OfS). 2022. 'Consultation on a New Approach to Regulating Student Outcomes'.
- Quintano, Claudio, Rosalia Castellano, and Antonella D'Agostino. 2008. "Graduates in Economics and Educational Mismatch: The Case Study of the University of Naples "Parthenope"." *Journal of Education and Work* 21 (3): 249–71. doi:10.1080/13639080802214118
- Schomburg, H. 2011. "Employability and Mobility of Bachelor Graduates: The Findings of Graduate Surveys in Ten European Countries on the Assessment of the Impact of the Bologna Reform." In *Employability and Mobility of Bachelor Graduates in Europe*, 253–73. Sense Publishing.
- Singer, Eleanor, John van Hoewyk, Nancy Gebler, Trivellore Raghunathan, and Katherine Mcgonagle. 1999. "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys." *Journal of Official Statistics* 15 (2): 217–30.
- Tomlinson, Michael. 2008. "'The Degree Is Not Enough': Students' Perceptions of the Role of Higher Education Credentials for Graduate Work and Employability." *British Journal of Sociology of Education* 29 (1): 49–61. doi:10.1080/01425690701737457
- Universities U. K. 2022. 'Higher Education in Numbers'. Insight and Analysis. 28 September 2022.