SRHE Society for Research into Higher Education

Governments harnessing the power of data to get “value for money”: Simulation studies evaluating England’s Office for Student B3 Progression metric

Research report

January 2024

Vladislav Areshka (Unviersity of Portsmouth) and Dr Alexander Bradley (University of Portsmouth)

UNIVERSITY OF PORTSMOUTH

Contents

Disclaimer: The views expressed in this report are the authors' and do not necessarily reflect those of the Society for Research into Higher Education

Acknowledgements

We are grateful to the Society for Research into Higher Education (SRHE) for funding this project.

We are also grateful to Professor Steven Jones for his advice as our critical friend and to Rob Gresham for his support throughout this study.

About the Authors

Vladislav Areshka is a PhD researcher in American literature at the University of Portsmouth. His research encompasses spheres of Anglo-American poetry, historical linguistics, stylistics, and digital humanities. He is currently developing a digital database of 19th-century New England pronunciation.

Dr Alexander Bradley is the Graduate Outcomes manager at the University of Portsmouth and formely a Senior Lecturer of Psychology. His diverse research interest range from interventions to improve employability, exploring how observability impacts altruistic acts to investigating how gambling companies utilise social media platforms. He is a fellow of the Higher Education Academy and can be contacted at alexander.bradley@port.ac.uk.

To cite this report: Areska, V & Bradley A. (2023). Governments harnessing the power of data to get “value for money”:Simulation studies evaluating England’s Office for Student B3 Progression metric. Society for Research into Higher Education (SRHE) Research Report.

Executive Summary

The regulator in England, the Office for Students (OfS), for higher education is there to ensure the quality of higher education provision both in terms of student experience and student outcomes. Our work focussed on the progression metric, which measures the extent to which graduates attain a positive outcome (i.e. a professional job, further study or another positive outcome) within 15 months of their course finishing. The progression metric is one of its key measures of assessing the quality of student outcomes through its regulatory monitoring, teaching excellence framework assessment (TEF) and Access and Participation Plans (APP). The OfS currently applies 90% and 95% confidence intervals to identify courses that are below its standards. Specifically, it uses a 90% confidence as a threshold for investiagtion where it could ask the university to provide more data on their outcomes and it could ask a team of investigators to inspect the providers' provision. The 95% confidence interval is a threshold for taking regulatory action such financial penalities or removing degree awarding powers if the findings from their investigation have not identified mitigating factors. Confidence intervals represent an estimated range from which the progression rate of the population will likely be. A 90% confidence interval means that there is a 90% confidence that the true population progression value will fall within that range. However, the use of 90% and 95% CI raises serious concern around a) how effectively it can accurately identify courses below standard (false positives) and b) how often does it identify courses as below standard when in fact they are at or above the threshold (false negatives).

The first study uses simulated data designed to explore the application of a 90% and 95% confidence interval as opposed to a 99% confidence interval in the OfS decision to investigate and take regulatory action in relation to false positives and false negatives. Our main findings were:

- The 90% confidence interval has a lower false positive rate than either the 95% CI or the 99% CI and is therefore a good choice as a threshold for investigation as it mitigates missing courses that are below standard.
- Larger sample sizes reduces false positives with sample sizes below 500 leading to higher rates of false positive which means every effort should be made to boost response rates to the Graduate Outcomes survey. Second, regulators should avoid regulating at micro-levels for example, small subject courses where sample sizes will be small.
- If the OfS used 99% confidence intervals as opposed to 95% to take regulatory action then false negatives (i.e. incorrect identification of a course as below threshold) would be reduced from around 2% of provision at threshold with 95% confidence interval to 0.2% with 99% confidence interval.

The second study uses simulated data designed to explore whether applying the Jeffries confidence interval, currently used by the OfS, for binomial distribution is better at reducing false positive and false negatives than other confidence intervals specialised for binomial distributions with small samples like Wilson and Agresti-Coull. The main findings were:

- The Wilson confidence interval performed better in terms of precision (narrow confidence interval), coverage probability (more likely to include population value within confidence interval) and had lower false positives but it did perform worse than Jeffries on false negatives. The differences in performances are marginal and unlikely to have real world consequences.

Recommendations, the OfS should keep using the 90% confidence interval as the threshold for investigation since the simulation shows this reduces false positives. The OfS should adopt the 99% Confidence interval over the 95% confidence interval as the threshold for regulatory action since the simulations shows this reduces false negatives. The use of both 90% and 99% CI gives the best protection against both false positive and false negatives. Based on the current OfS data over the last four years this is the difference of 10 false identified courses using the 95% CI compared to 1 false identified course(s) using the 99% CI.

Background

Increasing participation in higher education has led to greater spending by governments to fund institutions to deliver teaching and to students to facilitate their learning (Bondar et al., 2020; OECD, 2017). This coupled with a challenging period within the economy where world growth has been slow and inflation has been high (International Monetary Fund (IMF), 2023) has led to governments and their regulators looking to measure, evaluate and regulate universities to ensure good returns on their investments.

In England these wider structural forces have been have been met by England's higher education regulator the Office for Students (OfS) who have introduced a number of conditions on providers to maintain and improve the quality of higher education provision. One of the important conditions introduced is the B3 regulations which assess universities on the extent to which they support students to a) continue with their studies (continuation metric), b) complete their university course (completion metric) and c) to attain a positive outcome (i.e. professional career or further study) within 15 months of finishing their course (progression metric)(Office for Students, 2022a). The OfS have set different thresholds of quality that they deem acceptable for universities provision to achieve which varies by mode of study (i.e. full-time, part-time), level of study (undergraduate, postgraduate taught etc.) and split indicators like gender,ethnicity, disability, measures of deprivation etc (Office for Students, 2022b). For example, UK domiciled full time, first degree courses need to have 60% of their graduates with a positive outcome within 15 months of completing their studies to meet the threshold and avoid punitative regulatory action like fines or in the worst case scenario the revoking of degree awarding powers.

The challenge with the progression metric is that it is based on the Graduate Outcome survey, which has a response rate for full-time first degree graduates at around 56% (Higher Education Statistics Authority (HESA), 2020, 2022, 2023). Whilst this might be a reasonably good response rate to a large national survey it does mean that 44% of the population did not respond. This response rate introduces sampling error which is the extent to which the sample differs from the population across the items being measured (Banack et al., 2021). To quantify this uncertainty the OfS employs confidence intervals to allow them to ascertain the degree of certainty that a universitiy provider is above or below the progression thresholds. The OfS applies a series of confidence intervals from 75-99.7% however, they also state that they could potentially investigate a set of courses when the 90% confidence interval is below the required threshold and have strong statistical evidence if the 95% confidence interval is below the threshold. The challenges of setting thresholds and using confidence interval to identify those below, at or above standards is that you are likely to have false positives where one cannot correctly identify courses below thresholds and false negatives where one misidentifies that at or above threshold as below. As it currently stands, no evidence has been presented as to how

frequent false positive or false negatives are and how they are impacted by the use of a 90%, 95% confidence intervals as opposed to a 99% confidence interval.

Study one aimed to explore the consequences of regulatory choices in setting thresholds for investigating and taking enforcement measures against universities for poor graduate outcomes. Currently, the OfS uses a 90% confidence interval (CI) below a 60% threshold[1] to trigger an investigation and a 95% CI to provide evidence for taking regulatory action which can include financial penalties and even, at the extreme, revoking degree-awarding powers. The decision of the choice of a confidence interval is therefore an important consideration for the OfS and providers. Simulation of varying levels of confidence intervals are used to show how the choice of 90%, 95% or 99% confidence levels influences risks of misclassifying courses:

(a) which fall below a 60% threshold for positive outcomes, and

(b) incorrectly investigating and acting against courses at or above the threshold.

The second study aimed is to assess alternative methods of computing confidence intervals. Currently, the OfS uses the Jeffrey method (Office for Students, 2022d) However, research has shown that alternatives such as the Wilson and Agresti and Coull methods could be more precise in producing fewer false positives (Dean & Pagano, 2015; Franco et al., 2019). To date, there has been no empirical evidence produced by the OfS to support their choice in confidence intervals despite the fact we know different confidence intervals perform better under different conditions (i.e. Brown et al., 2001). Given the important implications that these calculations have for HEIs this is an omission that needs addressing. Simulations of Wilson, Agresti and Coull and Jeffries confidence interval are conducted using 90%, 95% and 99% so an empirical investigate of how different confidence interval perform in terms:

a) Capturing the true population value as accurately as possible
b) Correctly identifying courses below the threshold as below (false positive) and courses above the threshold as above (false negatives)

[1] As said above, threshold levels vary by mode and level of study. However, for readability the 60% threshold is chosen as this applies to first-degree, full time students who make up the majority of the student population.

Method

Both studies implemented a similar simulation protocol therefore we will describe the simulation process highlighting key differences between study one and two.

Simulation description

The simulation was designed to replicate the OfS B3 progression metric. In both simulations, we created three important variables.

First, the percentage of graduates in the population with a positive outcome varied from 20% up to 95% increasing 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 95% would be unlikely to be identified as below the threshold.

Second, we varied 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%, 35%, …, 90%)(Office for Students, 2022d).

Simulation procedure

Both simulations were created through the following three steps: data generation, sampling of the data, and calculating statistics from each of the samples. Only in the final step are their real differences between the simulations.

Data generation involved the creation of 97 datasets each with 16 columns (variables). The 97 datasets, representing student outcomes, went from 40 rows in length up to 1000 rows long in ten increments per dataset (e.g, the second dataset had 50 rows). Each row symbolises a graduate with a positive outcome (1) or negative outcome (0). Each column within a dataset represented a population with a certain percentage of positive outcomes. The first column in a dataset had 20% of the rows with a 1 symbolising a

population with 20% positive outcomes, whilst the second column had 25% with a positive outcome and this percentage increased by 5% in each column till the 16th column which had 95% with a positive outcome.

The second step was to take a randomly selected sample, with replacement, from each of the 16 columns 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 columns in all 97 datasets.

In the third step for study one in each of these samples we calculated four statistics. First is the percentage with a positive outcome/graduate job. Then three levels of confidence intervals – 90%, 95% and 99%. The Jeffries method has been used to calculate confidence intervals, since 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 et al., 2001; Office for Students, 2022d).

In the second study we also calculated the percentage with a positive outcome, in addition to the Jeffries method we used the Wilson and Acresti and Coull confidence intervals to calculate the 95% confidence intervals. Wilson and Acresti and Coull were calculated because other simulations have shown these methods have reasonable properties in binomial distribution with small samples (Dean & Pagano, 2015; Franco et al., 2019).

Simulations were performed with R Studio (version 2023.06.2 Build 561) running R version 4.3.0 using four packages ('tidyverse', 'DescTools', 'kableExtra', 'modelsummary'). All the data and code for simulation is available from the Open Science Framework (OSF).

Confidence interval measure

As a way of measuring performance of 90%, 95%, 99% confidence intervals, we evaluated their ability to correctly reflect a course's true population progression level as lying below or above the 60% progression threshold. Therefore, in analysing the choice of one confidence interval against another, two risk parameters were accounted for: 'false positive' and 'false negative' misclassifications. These two metrics were used in study one and two. For study two we also calculated two other statistics: confidence interval width

which is the range from the lowest confidence interval to the upper confidence interval (otherwise referred to as precision) and coverage probability, which is a more exact measure of the likelihood that the true population value lies within the lower and upper bounds of the confidence interval. Over the next two sections we outline in more details false positives and false negatives.

False positives

False positive misclassifications describe the cases when an upper limit of a confidence interval, that is calculated for an estimate from a population with less than 60% with a positive outcome, crosses the 60% threshold, and so there could be not enough statistical certainty to believe that a provider's underlying performance can indeed be below the threshold. For instance, Figure 1 shows different estimates for a population with 50% in positive outcomes and their respective confidence intervals set at 90% level of statistical confidence. It can be observed that in 72 out of 100 cases the upper limit of the confidence interval crosses the 60% threshold. Therefore, University 'X' could have an underperforming course with 50% of positive outcomes, yet – due to sampling and the quality of statistical evaluations around uncertainty levels – confidence intervals may fail to identify it correctly (or with strong enough statistical evidence) as below.

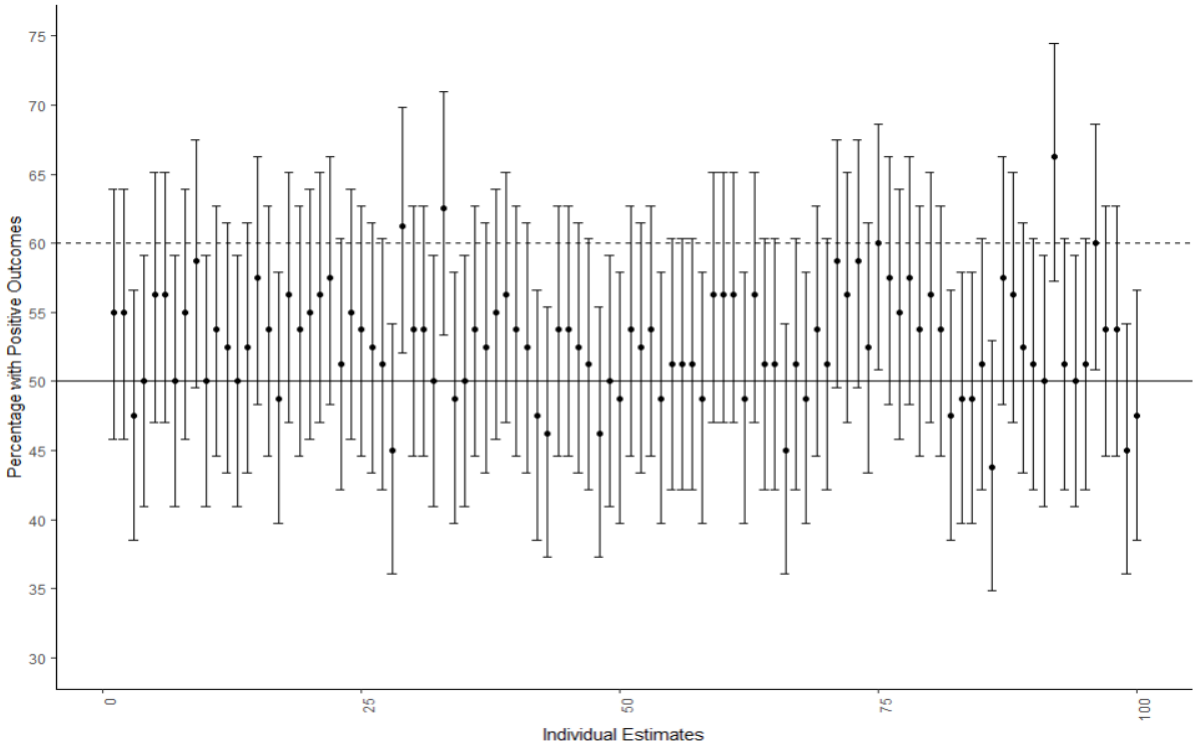

Figure 1. This graph depicts 100 estimates with 90% CI from a population with 50% of positive outcomes and a sample size of 80 students from a course of 160 students. The figure shows how many samples would not be properly identified as below the 60% threshold.

False negatives

False negative misclassifications denote the cases when a population level of positive outcomes is at or above the 60% threshold, but the sampled estimate and the upper limit of its respective confidence interval suggest the course is below the threshold. For example, Figure 2 has a population of 65% with positive outcomes and shows that 5 times out of 100 the samples with the upper limits of their 90% CIs falsely suggested that the underlying performance of the courses were below the threshold. Thus, University 'Y' could be wrongly identified as underperforming due to sampling when, in fact, the unknown level of positive outcomes was 65% – above the threshold.

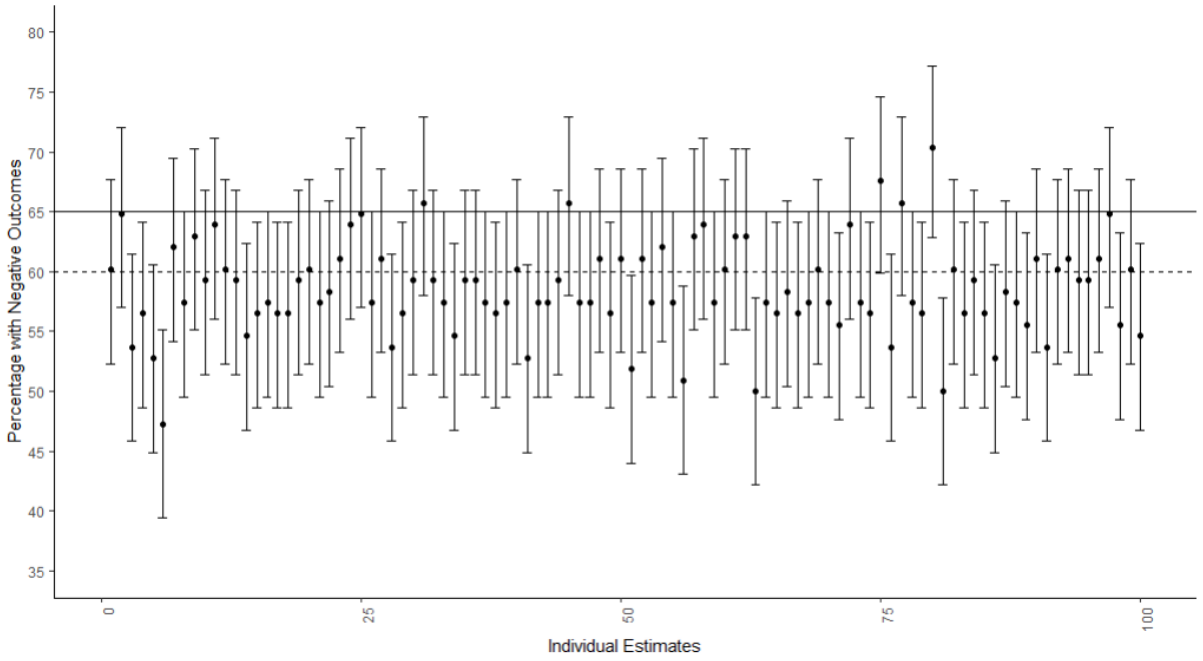

Figure 2. This graph depicts 100 estimates with 90% CI from a population with 65% of positive outcomes and a sample size of 108 students from a course of 360 students. Five samples would be identified as being below the 60% limit with upper confidence interval below the 60% threshold.

Results and Discussion Study One

Table 1 shows across all simulations, include all levels of sample size, percentage sampled, percentage in positive outcomes, the mean confidence interval range, number and percentage of false positives. The 90% confidence interval has smaller levels of false positives (8.22%) compared to either the 95% (10.59%) or 99% (15.71%) confidence intervals. This is in part due, to the varying range of the confidence intervals with the 90% confidence interval having a smaller average range between lower and upper bound of the confidence interval (9.76%) compared to either the 95% (11.61%) or 99% (15.18%) confidence interval.

Table 1.

Demonstrates the number and percentage of false positives by confidence intervals.

Confidence Interval	N° Total Estimations	Mean CI range	N° False Positives	% False Positives
90	998400	9.76	82115	8.22
95	998400	11.61	105777	10.59
99	998400	15.18	156804	15.71

Another factor that is worth considering in relation to false positive is how it changes as the percentage in positive outcomes approaches the threshold because a sample with 55% in positive outcomes ought to have a much large false positive rate than a sample were only 40% have a positive outcome which is much further away from the 60% threshold. Table 2 clearly demonstrates that as the percentage in graduate jobs (positive outcomes) increase so to does the percentage of false positive (i.e. those courses that cannot be identified as below the threshold). While, the 90% confidence interval performs better at discerning those courses of provision below the threshold than 95% or 99% however, with 55% in graduate jobs it has a false positive rate of 45.13%, 95% has a false positive rate of 56.27% and 99% has a false positive rate of 76.41%. The implication of this is the OfS will finding it difficult to correctly identify with statistical certainty courses close below the threshold the closer they get to the 60% threshold.

Table 2.

Illustrates how the percentages in graduate jobs impact the number of false positives for 90%, 95% and 99% CIs.

% Population in Graduate Job	N° Total Estimations	N° False Positives 99CI	% False Positives 99CI	N° False Positives 95CI	% False Positives 95CI	N° False Positives 90CI	% False Positives 90CI
20	124800	30	0.02	4	0.00	1	0.00
25	124800	69	0.06	15	0.01	4	0.00

30	124800	765	0.61	224	0.18	109	0.09
35	124800	1840	1.47	564	0.45	282	0.23
40	124800	8028	6.43	3972	3.18	2456	1.97
45	124800	15788	12.65	9712	7.78	7073	5.67
50	124800	34920	27.98	21060	16.88	15864	12.71
55	124800	95364	76.41	70226	56.27	56326	45.13

Another important factor that is likely to impact false positive rates is the sample size since we know that smaller sample will have less precise estimates (large range of confidence interval) we can therefore expect higher rates of false positives in smaller samples. Figure 3 does indeed show that false positive rates are higher in smaller sample sizes and especially high as the percentage in graduate jobs approaches the threshold. Sample sizes of only 51-100 are required when 40% have a positive graduate job to get the false positive rate for the 90% confidence interval below 10%. However, at 45% in positive graduate jobs a sample of 101-150 is required to get the 90% confidence interval below 10% false positive whilst for 50% in positive graduate jobs a sample of 151-200 is need for the 90% confidence interval. At 55% in positive graduate jobs a sample of 651-700 is need to reduce false positive to less than 10% using the 90% confidence interval. The important finding for the OfS is that large sample sizes are required if they are to be able to accurately identify courses that are just below threshold. This has two key implications: first, regulators should avoid trying to apply micro-regulation at course level or any unit where sample sizes are likely to be very small since they will not with a good degree of statistical accuracy be able to identify those course below the threshold, especially just below. Second, both the OfS and the Higher Education Statistics Authority (HESA) who is responsible for conducting the graduate outcomes survey need to work together with providers to ensure high response rates to the graduate outcome survey. Lower response rates will mean smaller samples which in turn will impact the OfS ability to accurately identify courses below thresholds.

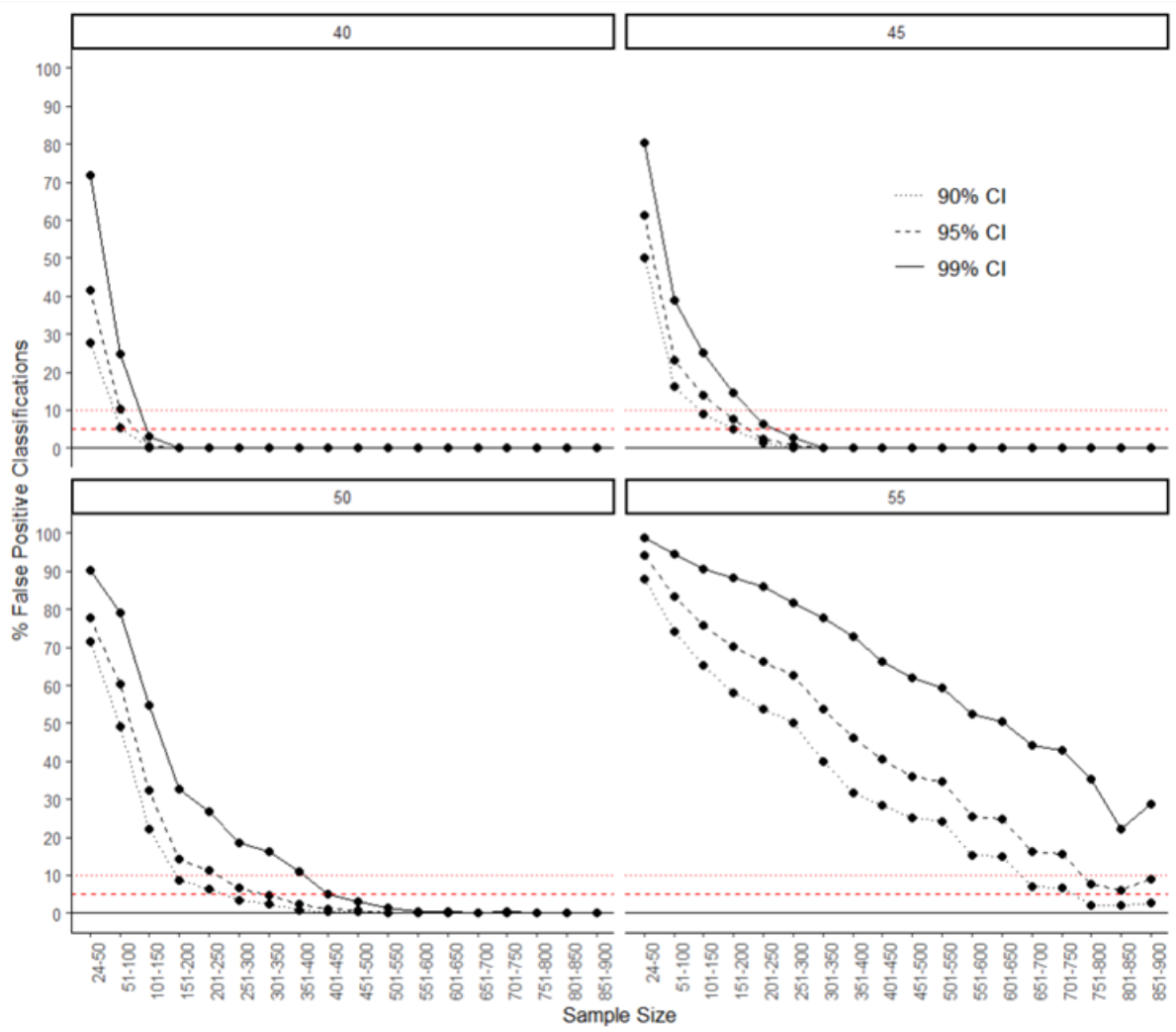

Figure 3. These graphs depict the percentage of false positive cases by confidence interval and sample size that is sampled from the population. Each graph represents the percentage with a positive outcome in the population starting from 40% top left to 55% bottom right. The horizontal lines represent levels of error expected for each confidence interval: the dotted line (•••) – 10% error for 90% CI; the dashed line (- - -) – 5% error for 95% CI.

The 90% confidence interval performs well on false positive it performs less well than the 95% and 99% confidence interval (see Table 3) on false negative. For example, the 90% confidence has double the rate of false negatives than the 95% (0.57% vs 0.24%) confidence interval which itself has far higher rate false negative rates than the 99% confidence interval (0.03%). Table 3 is an overly flattering display on false negatives since we would not expect samples with 70 or 80% in graduate outcomes to be incorrectly identified as below the threshold. We now explore how the percentage in graduate outcomes and the sample size impacts false negatives across 90%, 95% and 99% confidence intervals.

Table 3.

Demonstrates the number and percentage of false negatives by confidence intervals on samples with 60% or above in positive outcomes.

Confidence Interval	N° Total Estimations	Mean CI range	N° False Negatives	% False Negatives
90	998400	9.76	5656	0.57
95	998400	11.61	2372	0.24
99	998400	15.18	258	0.03

Table 4 shows how false negative rates are higher at the threshold and negligible at 65% in graduate outcomes and beyond. At the 60% threshold we can see that false negative rates are over twice as high using 90% confidence interval compared to a 95% interval which is itself much higher than the 99% confidence interval. To assess the practical impact of these difference we downloaded the OfS data for the progression metric based on 2017-2020 surveys of courses at all levels and modes of study (excluding those with no outcome indicator values). We found that there were 534 cases where the sample estimate was at or within one percent above the threshold. As it stands the OfS would use the 95% confidence interval which would incorrectly identify 10 cases as below standard which in fact would be above at or above the threshold. If the 99% confidence interval was used instead that would only be 1 case. It should be noted that in between identification and financial/regulatory penalties the OfS does conduct an investigation with inspectors on the ground (for example, see Office for Students 2023), nevertheless the procedures this follows is yet to be fully spelt out. It would seem prudent for the OfS to adopt 99% CI instead of 95% CI to provide greater confidence and trust within its regulatory framework since it can't be sure which area of provision that is below the threshold with 95% CI is below due to sample error, as opposed to their real population value falling short of the threshold.

Table 4.

Illustrates how the number of false negatives fluctuates for 90%, 95% and 99% CIs depending on the percentages in graduate jobs.

% Population in Graduate Job	N° Total Estimations	N° False Negatives 99CI	% False Negatives 99CI	N° False Negatives 95CI	% False Negatives 95CI	N° False Negatives 90CI	% False Negatives 90CI
60	124800	255	0.2	2327	1.86	5502	4.41
65	124800	3	0.0	43	0.03	148	0.12
70	124800	0	0.0	2	0.00	6	0.00
75	124800	0	0.0	0	0.00	0	0.00
80	124800	0	0.0	0	0.00	0	0.00
85	124800	0	0.0	0	0.00	0	0.00
90	124800	0	0.0	0	0.00	0	0.00
95	124800	0	0.0	0	0.00	0	0.00

Finally, Figure 4 shows that there is no clear pattern between sample size and false negatives and it also reinforces the idea that the probability of false negatives is reduced with the adoption of the 99% confidence interval compared to 95% and 90% intervals.

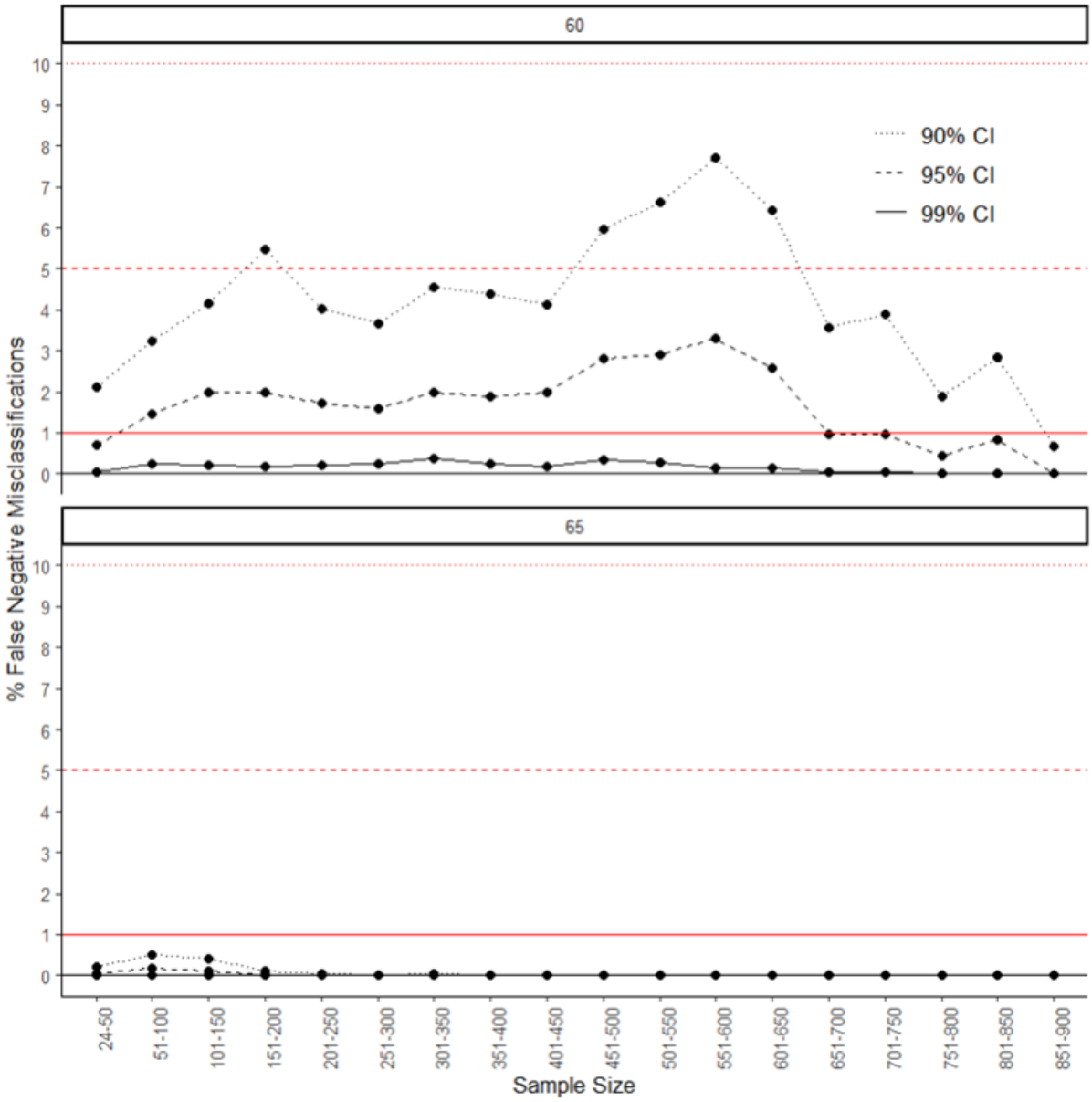

Figure 4. These graphs illustrate the percentage of false negative cases by confidence interval and sample size that is sampled from the population. The two graphs represent the percentage with a positive outcome/graduate job in the population: 60% top and 65% bottom. The horizontal lines represent levels of error expected for each confidence interval: the dotted line (•••) – 10% error for 90% CI; the dashed line (- - -) – 5% error for 95% CI; the straight line (—) – 1% error for 99% CI.

Results and Discussion Study Two

Table 5 illustrates that the Wilson interval on average across all simulations seemed to have marginally better precision (9.75% 90 CI; 11.59% 95CI; 15.12% 99CI) than both the Jeffries (9.76% 90 CI; 11.61% 95CI; 15.18% 99CI) and Agresti-Coull (9.80% 90 CI; 11.66% 95CI; 15.27% 99CI) method of calculating confidence interval (i.e. small range between upper and lower bounds of the confidence interval). Jeffries performs marginally worse on coverage probability than either Wilson or Agresti-Coull both of which were able to capture the true population value within their confidence interval more times than Jeffries; however, the differences are marginal.

Table 5.

Average ranges and coverages by 90%, 95%, 99% confidence levels.

Type of CI	Average range of 90% CI	Average range of 95% CI	Average range of 99% CI	Average Coverage of 90% CI	Average Coverage of 95% CI	Average Coverage of 99% CI
AC	9.80	11.66	15.27	90.07	95.09	99.03
J	9.76	11.61	15.18	89.85	94.90	98.96
W	9.75	11.59	15.13	89.91	94.97	98.97

Figure 6 illustrates how sample size impacts both coverage probability and precision (relative width of confidence intervals). We can see that the Wilson interval compared to Jeffries performs slightly better in terms of precision and marginally higher coverage probability in samples less than 200 across the 90%, 95% and 99% confidence interval. Agresti-Coull often performs better with higher coverage probability than Jeffries; however this comes at the cost of slight worse precision (i.e. large range between lower and upper bound of CI). Similarly, once samples approach 200 then the choice of confidence interval does not seem to impact coverage probability or precision.

The choice of confidence interval has minimal impact on false positive rates and certainly not to the extent of having real world consequences (see Table 6). For example, at 55% with a graduate job there was less than a 1% difference in false positive rates between Jeffries, Wilson and Agresti-Coull confidence intervals. A similar picture emerges when we look at false negatives when we look across different levels of percentage in positive graduate outcomes (see Table 7). Similar to study one we see the largest false negative rates at the threshold and negligible rates at 5% or more above the threshold. At the 60% threshold the Jeffries method has 1.88% with false negatives compared to the Wilson and Agresti-Coull which have slightly higher false negative rates at 1.95%. This 0.07% difference is unlikely to have real world implications. On balance, whilst the Wilson has marginally better precision, coverage probability in small samples and lower false positives it performs worse on false negatives. Furthermore, these differences are marginal and are unlikely to have real world implications.

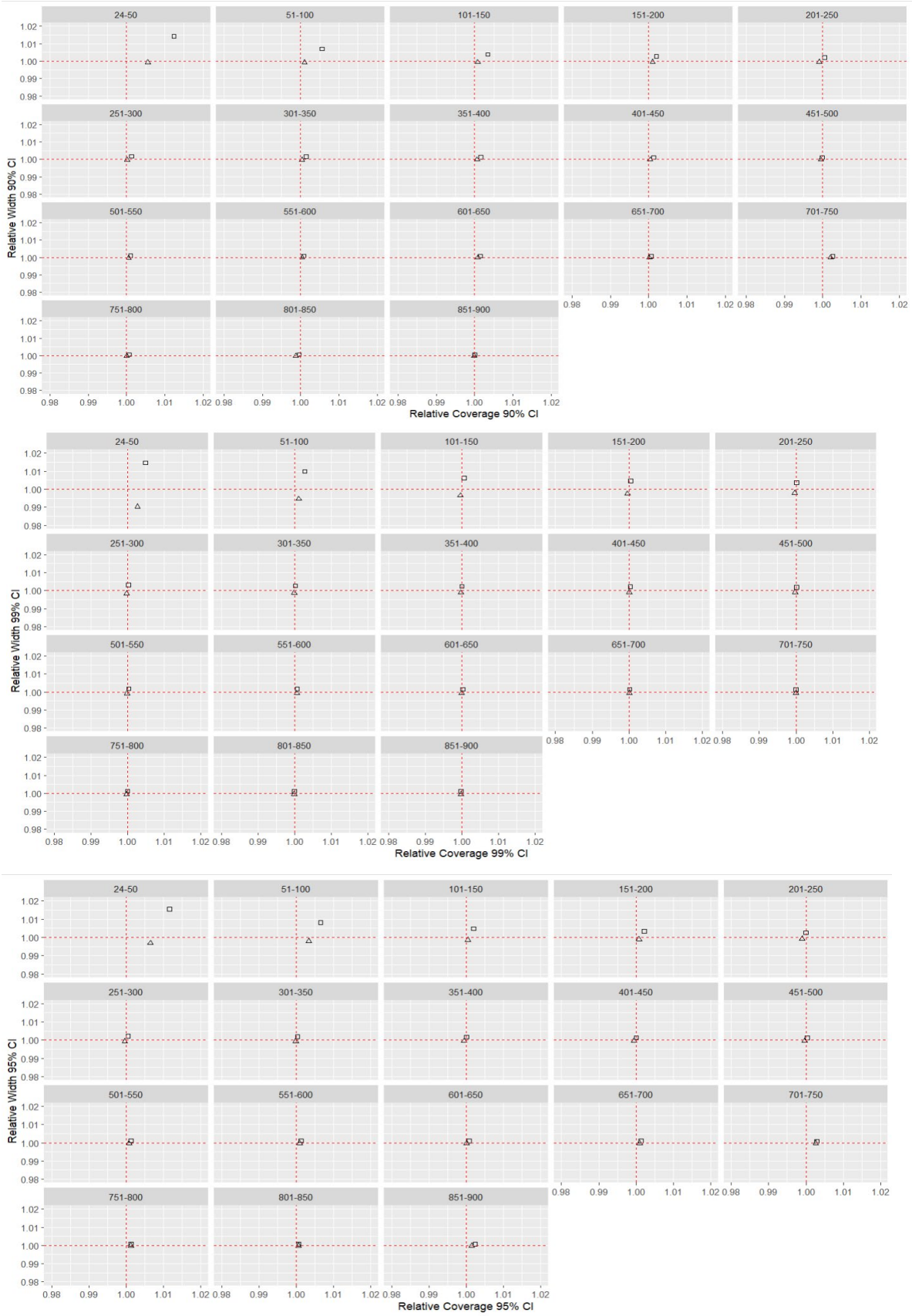

Figure 5. Relative performance of Agresti-Coull (□) and Wilson (△) CIs against Jeffreys (red dashed lines at 1 point) at 90%, 95%, 99% confidence levels. The best-case scenario – a CI indicator falling in a lower right square.

Table 6

Performance by False Positives at 95% confidence level across Agresti-Coull, Jeffries and Wilson

% Population in Graduate Job	N	Count UCI AC 95 Less 60	% UCI AC 95 Less 60	Count UCI 95 J Less 60	% UCI J 95 Less 60	Count UCI 95 W Less 60	% UCI 95 Less W 60
20	124800	8	0.01	8	0.01	8	0.01
25	124800	13	0.01	13	0.01	13	0.01
30	124800	189	0.15	189	0.15	189	0.15
35	124800	565	0.45	573	0.46	565	0.45
40	124800	3888	3.12	3967	3.18	3888	3.12
45	124800	9564	7.66	9720	7.79	9564	7.66
50	124800	20922	16.76	21121	16.92	20922	16.76
55	124800	69914	56.02	70370	56.39	69914	56.02

Table 7

Performance by False Negatives across Agresti-Coull, Jeffries and Wilson 95 Confidence interval

% Population in Graduate Job	N° Total Estimations	N° False Negatives AC	% False Negatives AC	N° False Negatives J	% False Negatives J	N° False Negatives W	% False Negatives W
60	124800	2433	1.95	2349	1.88	2433	1.95
65	124800	59	0.05	58	0.05	59	0.05
70	124800	1	0.00	1	0.00	1	0.00
75	124800	0	0.00	0	0.00	0	0.00
80	124800	0	0.00	0	0.00	0	0.00
85	124800	0	0.00	0	0.00	0	0.00
90	124800	0	0.00	0	0.00	0	0.00
95	124800	0	0.00	0	0.00	0	0.00

Project Outputs to Date

The findings from study one have been presented at the SRHE Higher Education Conference in 2023.

-Areshka, V and Bradley, A. 2023. What does simulations of the Office for Students B3 regulations tell us about how fair and effectively it can identify courses below specified thresholds. Society for Research into Higher Education Conference: Higher Education Research, Practice, and Policy: Connections & Complexities. December 2023. Online.

The findings from study one are currently under review at the journal: Studies in Higher Education.

-Areshka, V and Bradley, A. 2024. Is England Office for Students likely to falsely identify courses as below threshold on the B3 Progression Metric? Studies in Higher Education. (submitted, under review)

We aim to write up these findings for a non-academic audience in potential HE relevant websites like SRHE blog and WonkHe.

Recommendations

The results of the simulations lead the authors to make the following recommendations.

For the Office for Students (OfS):

1) The 99% confidence interval should be used as the threshold of strong statistical evidence as this would mitigate risks of false negatives compared to the current use of the 95% Confidence interval. However, the 90% confidence interval should still be used as a threshold for investigation since this mitigates against false positives. Thus combining the 90% with the 99% CI would give you the best protection against false positive and false negatives.
2) When planning future regulations every effort should be made to avoid creating regulation that would be applied to small samples since this will lead to far higher false positive rates and in effect an inability to correctly hold those below standard to account.
3) Every effort should be made to work with HESA and providers to boast response rates to the Graduate Outcomes survey to mitigate the risks of false positives.

For HESA and Providers:

1) Every effort should be made to increase response rates to the graduate outcomes survey since lower response rates will lead to smaller samples and ultimately, higher false positives.

References

Banack, H. R., Hayes-Larson, E., & Mayeda, E. R. (2021). Monte Carlo Simulation Approaches for Quantitative Bias Analysis: A Tutorial. In *Epidemiologic Reviews* (Vol. 43, Issue 1, pp. 106–117). Oxford University Press. https://doi.org/10.1093/epirev/mxab012

Bondar, T. I., Telychko, N. V., Tovkanets, H. V., Shcherban, T. D., & Kobal, V. I. (2020). Trends in Higher Education in EU Countries and non-EU Countries: Comparative Analysis. *Revista Romaneasca Pentru Educatie Multidimensionala*, *12*(1Sup1), 77–92. https://doi.org/10.18662/rrem/12.1sup1/224

Brown, L., Cai, T., & Dasgupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, *16*(2), 101–133.

Dean, N., & Pagano, M. (2015). Evaluating confidence interval methods for binomial proportions in clustered surveys. *Journal of Survey Statistics and Methodology*, *3*(4), 484–503. https://doi.org/10.1093/jssam/smv024

Franco, C., Little, R. J. A., Louis, T. A., & Slud, E. V. (2019). Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys. *Journal of Survey Statistics and Methodology*, *7*(3), 334–364. https://doi.org/10.1093/jssam/smy019

Higher Education Statistics Authority (HESA). (2020). *Higher Education Graduate Outcomes Statistics: UK, 2017/18*. Graduate Outcomes Survey.

Higher Education Statistics Authority (HESA). (2022, June 16). *Graduate Outcomes 2019/20: Summary Statistics - Summary*. Summary Statistics Graduate Outcomes.

Higher Education Statistics Authority (HESA). (2023). *Graduate Outcomes 2020/21: Summary Statistics - Summary*.

International Monetary Fund (IMF). (2023). *World economic outlook: Navigating a global divergence.* www.bookstore.imf.org

OECD. (2017). *Benchmarking higher education system performance: Conceptual framework and data, Enhancing Higher Education System Performance.*

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. (2023). *Quality Assessments*. Assessment Reports.