Teaching Excellence and Student Outcomes Framework (TEF): Findings from the subject-level pilot 2018-19

Annex H: Type I and Type II errors in the TEF metric flags

This report was completed in autumn 2019 following the conclusion of the pilot.

Publication date 21 January 2021

Contents

Introduction

1. This annex reports on analysis that the OfS has undertaken, since the subject-level pilot concluded, to describe the relationship between cohort size and how flags are currently generated. A general understanding of the TEF metrics and the associated flagging approach is assumed throughout[1]. The analysis was undertaken in response to questions that arose in the pilot about the cohort sizes needed to produce robust metrics at subject level.

2. The analysis looks at how far the likelihood of differences in performance being correctly flagged depends on the cohort size, in the current method. The null hypothesis in the TEF is that the difference between the unit indicator and benchmark is zero. In practical terms this means that the unit's performance is no different to how the sector as a whole is performing. In this case, the unit refers to a provider or a provider's subject.

    a. A Type I error is the rejection of a true null hypothesis. Usually a type I error leads to the conclusion that a supposed effect or relationship exists when in fact it does not. In TEF, this would be the flagging of a unit when, in reality, the performance of the unit is no different to the benchmark (or sector adjusted average).

    b. A Type II error is the failure to reject a false null hypothesis. In TEF, this would be the failure to flag a unit that is known to have a performance different to the sector.

3. Type I and Type II errors are related, and typically, reducing the likelihood of one might be expected to increase the likelihood of the other. It follows that policy makers and other users of statistics will want to balance the likelihoods of each type of error appropriately to their use in context when determining the parameters to be applied in their interpretation of the statistics.

4. To understand the extent of Type I and Type II errors in the current TEF method, we have applied a statistical simulation approach (described in paragraphs 7 to 11) which tests artificial adjustments to the performance of a unit for a given metric, and shows the likelihood that this adjustment in performance results in an incorrect interpretation of the metric via the existing flagging method. For clarity and ease of interpretation, the analysis is repeated here for the full-time student cohorts across three different metrics (NSS scale 1 - 'teaching on my course', continuation and highly-skilled employment), and focuses on the results observed in subject-level metrics as the unit of reference. However, we have no reason to believe that the findings reported here do not generalise to the complete set of TEF metrics, to metrics generated for part-time student cohorts, and to metrics which consider the provider as the unit of reference.

Simplified Type II error calculation

5. Using the normal approximation of the binomial distribution, we can calculate the size of the 95 per cent confidence interval for the absolute difference between an observed indicator (for example, the observed continuation rate) and its associated benchmark. Assuming there is an actual underlying difference, if this confidence interval contains the value of the observed difference that we are trying to detect, there will be a high level of Type II errors. That is, there is

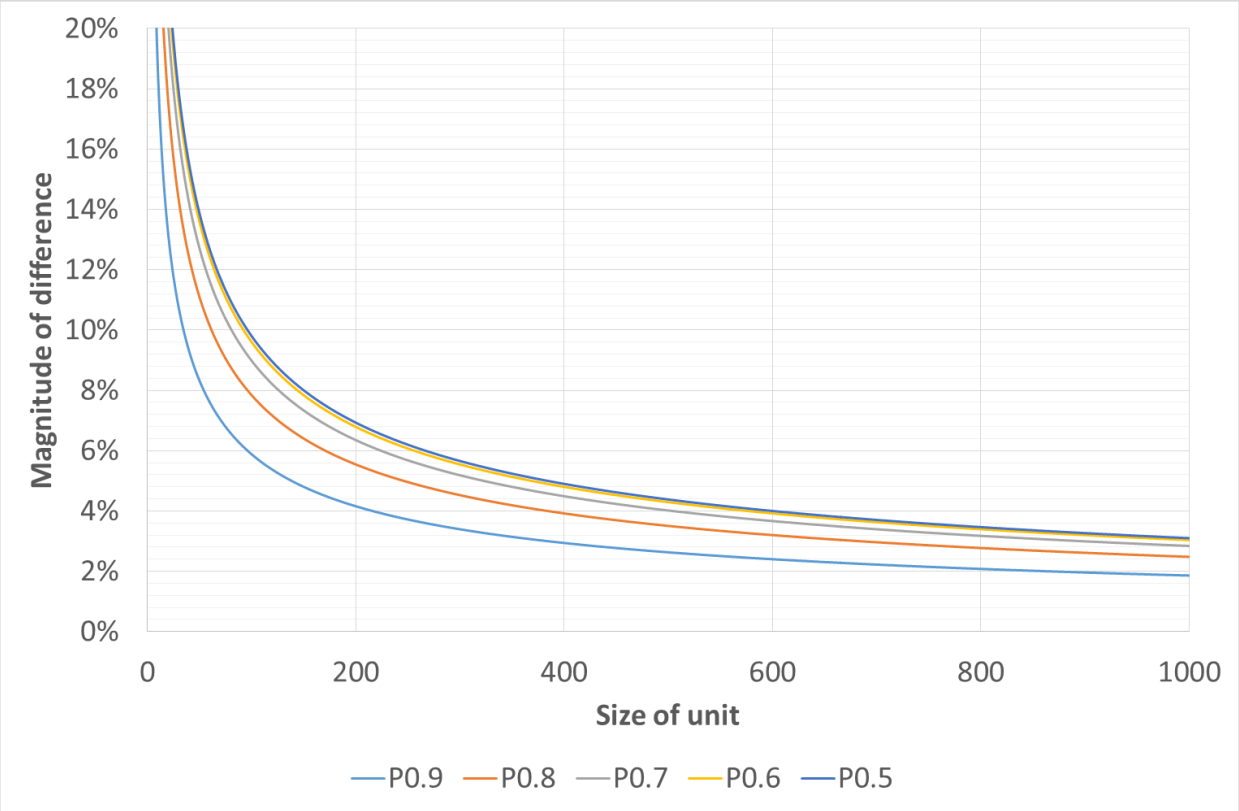
[1] For further information about the current TEF methods, see the specification available at www.gov.uk/government/collections/teaching-excellence-framework.

a strong chance that actual, non-zero differences are not flagged. The size of this confidence interval varies depending on the unit size and the indicator (when expressed as a proportion).

6. Figure 1 shows how this confidence interval varies by size of unit. The lines represent different proportions for the indicator, for example the P0.9 line represents an indicator of 90 per cent.

7. Figure 1 suggests that for any indicator value, in order to reliably detect absolute differences of six or more percentage points, the size of the unit needs to be in excess of 200. When differences of two or more percentage points are under scrutiny, the size of unit needs to be over 800.

Figure 1: Normal approximation to the binomial



8. In order to gain a simplified approximation of what magnitude of percentage point difference between the unit indicator and benchmark would be required to normally be detected and flagged, we can focus on the uncertainty in the unit indicator. This is a flawed assumption as there is uncertainty in the benchmark as well. However, it should still provide a good indication of what the minimum unit size cannot fall below for robust detection (i.e. minimising the Type II error). This is described in the simulation methodology below.

Simulation methodology

9. A Bayesian simulation approach is used to estimate the level of Type I and Type II errors in the current TEF methodology for identifying units that are performing significantly different to their benchmark. This calibration is applied to a single core metric. In this annex, the initial metric examined is National Student Survey (NSS) Scale 1 (‘teaching on my course’) and the units examined are subject level within provider. The size of units and distribution of student groups is as observed in the actual TEF data.

10. The methodology for generating the simulations are as follows:

    a. Fit a fixed effects model using the TEF individualised student-level data with NSS Scale 1 as the outcome of interest (dependent variable), and the benchmarking student characteristics as main effects (independent variable).

    b. Calculate the expected probability of a positive NSS Scale 1 outcome for each individual student using the fixed effects model. The original outcomes should be used to calculate these probabilities. Note that the probability for each individual is unaffected by the unit they are recorded within.

    c. For each individual in the dataset, simulate whether the individual has a positive NSS Scale 1 outcome.

    d. Calculate the TEF unit level flags using the standard TEF methodology but with the simulated individual outcomes. Record the proportion of units flagged and not flagged, where single and double flags are treated equally as flagged.

    e. Repeat steps c. and d. until convergence is achieved[2] for the statistics being monitored - in this case, the proportion of units flagged and not flagged.

11. This methodology allows for the assessment of Type I errors as the simulated data is designed so that all units perform in the same way. Therefore, any units where the null hypothesis is rejected (or where flagging occurs) are incorrect.

12. This simulation methodology can be modified to also produce estimates of Type II errors by adjusting the probabilities in the same way for all students in a small set of randomly selected units. These adjusted units are then monitored separately. The flagging of these units would be expected as these units are designed to be operating differently to the sector. Any failure to flag these units within the methodology would be seen as a Type II error.

13. The selection of these units and adjustment of probabilities are carried out as follows.

    I. For each simulation run (before step b.), stratify the units into size categories:

    - Less than 20 individuals in the unit
    - 20 to 49 individuals
    - 50 to 99 individuals
    - 100 to 249 individuals
    - 250 to 499 individuals
    - 500 to 999 individuals
    - 1,000 or more individuals.

    II. For each of these size categories, randomly select four units from each size category. For two of the units, add X percentage points[3] to each individual's expected probability within the unit. For the other two selected units, subtract X percentage points for each

[2] For the simulations carried out in this annex, 50 simulations were undertaken and achieved convergence.

[3] In the tables, PPT2 references the detection of a 2 percentage point difference. PPTX references the detection of a X percentage point difference.

individual's expected probability. X will vary depending on the magnitude of the difference that the methodology is expected to detect. Only a small number (four in this case) of units' probabilities are modified as modification of all, or the majority, of units would mean the sector average would also be changed and the comparisons invalid.

III. Use the unadjusted and adjusted probabilities to simulate outcomes for each individual (step c.), and repeat the appropriate TEF flagging calculation (step d.).

IV. Repeat steps II-III until convergence is achieved.

Results for NSS Scale 1, subject units

14. Table 1 shows the proportion of units, by size, that have an incorrect flag (either positive or negative) – a Type I error. The error rates seen are close to the expected and designed error rate of 5 percentage points showing that the methodology is well-calibrated with regard to Type I errors.

Table 1: Type I errors in TEF flagging, NSS Scale 1, subjects as units

Size of unit	PPT2	PPT4	PPT6	PPT8	PPT10
Below 20	4%	4%	4%	4%	4%
20 to 49	4%	4%	4%	4%	4%
50 to 99	3%	3%	3%	3%	3%
100 to 249	3%	3%	3%	3%	3%
250 to 499	3%	3%	3%	3%	3%
500 to 999	3%	3%	3%	3%	3%
Above 1,000	4%	4%	4%	4%	4%

15. Table 2 shows the proportion of units, by size, with a designed difference that are not flagged using the TEF methodology (i.e. a Type II error). The effect of changing the magnitude of the designed difference has also been examined: designed unit differences from between two and 10 percentage points are reported.

16. The results show that for very small units, there is a very high rate of Type II errors regardless of the magnitude of the difference being assessed. For large units and differences large in magnitude, Type II errors are small in nature. For units in the middle size categories, the level of Type II error varies depending on the magnitude of the difference being assessed.

17. In much of the statistical literature, the acceptable level of Type II errors is higher (at around 20 per cent) than Type I (where five per cent is often used). These results indicate that for detecting differences of six percentage points or more, you would need unit sizes in excess of 250. For detecting differences of two percentage points or more, these unit sizes would need to be much more than 1,000: a 45 percentage Type II error is well in excess of the 20 per cent acceptable level. These results are in line with the simplified Type II calculations described earlier, as the uncertainty in the benchmark was not recognised in the simplified calculations.

Table 2: Type II errors in TEF flagging, NSS Scale 1, subjects as units

Size of unit	PPT2	PPT4	PPT6	PPT8	PPT10
Below 20	96%	98%	96%	96%	96%
20 to 49	98%	94%	82%	63%	42%
50 to 99	91%	85%	74%	57%	34%
100 to 249	90%	68%	39%	16%	5%
250 to 499	79%	36%	11%	3%	0%
500 to 999	79%	22%	2%	0%	0%
Above 1,000	45%	2%	0%	0%	0%

Other results

18. We have examined the Type II errors under other TEF conditions. While the extent varies slightly, the results show an apparent dependency on cohort size across all of the conditions considered here. When the cohort size is smaller than around 500 students, the likelihood of Type II errors is generally higher.

Table 3: Type II errors in TEF flagging, NSS Scale 1, providers as units

Size of unit	PPT2	PPT4	PPT6	PPT8	PPT10
Below 20	94%	96%	95%	92%	91%
20 to 49	96%	90%	82%	60%	46%
50 to 99	93%	87%	73%	54%	34%
100 to 249	87%	67%	36%	17%	6%
250 to 499	85%	48%	13%	3%	1%
500 to 999	76%	24%	3%	0%	0%
1,000 to 2,499	32%	1%	0%	0%	0%
2,500 to 4,999	3%	0%	0%	0%	0%
5,000 to 9,999	2%	0%	0%	0%	0%
10,000 and above	0%	0%	0%	0%	0%

Table 4: Type II errors in TEF flagging, continuation metric for full-time entrants, subjects as units

Size of unit	PPT2	PPT4	PPT6	PPT8	PPT10
Below 20	92%	92%	90%	89%	88%
20 to 49	88%	83%	72%	59%	49%
50 to 99	90%	82%	66%	47%	31%
100 to 249	79%	51%	23%	10%	5%
250 to 499	70%	16%	4%	1%	1%
500 to 999	54%	19%	13%	11%	10%
1,000 and above	18%	0%	0%	0%	0%

Note: Full convergence not met for some larger units.

Table 5: Type II errors in TEF flagging, highly skilled employment metric for graduates from full-time provision, subjects as units

Size of unit	PPT2	PPT4	PPT6	PPT8	PPT10
Below 20	94%	95%	95%	91%	84%
20 to 49	94%	91%	86%	78%	68%
50 to 99	95%	91%	81%	70%	54%
100 to 249	96%	84%	60%	30%	12%
250 to 499	87%	65%	42%	32%	27%
500 to 999	86%	50%	22%	13%	11%
1,000 and above	56%	2%	0%	0%	0%

Note: Full convergence not met for some larger units.