ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

RSS RESPONSE TO OFFICE FOR STUDENTS'S CONSULTATION ON THE TEACHING EXCELLENCE FRAMEWORK

17 March 2022

1. Summary

The RSS has been actively engaging with the Teaching Excellence Framework (TEF) since 2016. Our views on the scheme as it has been developing are outlined below:

- Response to Year Two Consultation (2016)
- Response to the Subject-Level Consultation (2018)
- Submission to the Independent Review of the TEF (2019)
- Letter to Ed Humpherson re: Teaching Excellence and Student Outcomes Framework (2019)

Throughout, we have primarily sought to identify and explain statistical problems with TEF so that the process can be improved as much as possible – as this is where we feel that the Royal Statistical Society brings an important perspective. Our main concern has been that the data-based indicators which are intended to measure teaching excellence are unlikely to actually do so.

The current proposals for TEF represent an improvement on previous versions, particularly with respect to the communication of uncertainty associated with the data-based indicators, but we continue to have some concerns, which we set out in this response. Our main concerns are summarised here:

1. Some of the data-based indicators may still incentivise behaviours that may not necessarily enhance teaching excellence.
2. Given the limitations of the data-based indicators, it is welcome that there is a greater emphasis on expert review – but this increases the importance of transparent panel guidance to help ensure that ratings are reproducible.
3. There is insufficient attention paid to the issue of multiple comparisons – and this problem is heightened when indicator data are presented on a huge number of splits.

2. Does TEF effectively incentivise excellence in teaching?

The stated purpose of the TEF is to "incentivise excellence in teaching, learning and student outcomes. The TEF should incentivise a provider to improve and to deliver excellence above our minimum baseline quality requirements, for its mix of students and courses. " (TEF consultation, p.12).

The consultation sets out a series of data-based indicators that will inform TEF ratings, relating to continuation (the percentage of students continuing study after their degree), completion (the percentage of students that complete their degree), progression (measured via Graduate Outcomes survey data) and student experience (measured by the National Student Survey).

It is acknowledged widely, including in the consultation document, that these indicators are, mostly, only indirectly associated with quality of teaching, learning and student outcomes. This is particularly true when considering some of the detail of how the indicators are counted: eg, any level of further study will be counted as a positive outcome, even if it is at a lower level than the degree that the student completed.

The RSS has been highly critical of past versions of TEF. In our submission to the independent review into TEF (p.6) we argued that the metrics previously used likely did not measure teaching quality: we suggested that research was required to assess whether the indicators were effective measures of teaching quality.

While the new proposals are an improvement on previous generations – in that the data deficiency is more explicitly acknowledged and the role of expert judgement is given an enhanced status – there remains a risk that TEF continues to provide an incentive to providers to focus their efforts on merely optimising the indicators. It is conceivable that some of the indicators might be optimised at the expense of more directly optimising excellence in teaching and learning. In this respect the very existence of TEF may be counterproductive to its aims.

What follows is therefore conditional on the fact that TEF exists and where we express agreement with a proposal should not be interpreted as supporting the TEF concept more broadly.

3. The four-year cycle of ratings at provider-level only

We are pleased to see that subject-level TEF ratings, of which we were highly critical in previous consultations (see our response to the 2018 consultation), are no longer planned. Our concerns about the robustness of the indicator data at subject level was also raised in the independent review and, particularly, in the ONS study that informed that review.

We recognise that there are some advantages to the proposed four-year cycle: especially in that it allows some smoothing of year-on-year variation in the data indicators. However it does also mean that TEF awards will be based on data from up to seven years previously. We do not have an opinion on whether four years provides the best balance of these two aspects while also not imposing an excessive burden on the sector – but think it is important to stress that there is a balance that here that needs to be consciously struck.

4. Assessing excellence in student experience and student outcomes

As we have raised in previous consultations, university educational experience and student outcomes is a complex multi-dimensional process. We consider TEF as over-simplistic in attempting to form a single judgement at provider level. Providing judgments on two aspects is preferable, but we note that the independent review proposed four aspects – teaching and learning environment, student satisfaction, educational gains and graduate outcomes. Having assessed a provider on multiple aspects it seems unnecessary and simplistic to then combine the aspect ratings into a single overall rating.

We also have some concerns on the specific plans outlined in the document. In particular we believe that the student outcomes aspect is being defined too narrowly. So, for example, issues of the challenge and currency of curriculum and students skill development which appear under student experience in Table 3 would be much more appropriate as items to be assessed under student outcomes.

5. Gold/Silver/Bronze ratings

We think it is important to note that the independent review strongly argued against Gold/Silver/Bronze names for TEF ratings. While the intention might be that gold, silver and bronze all indicate some level of success, the perception is clearly that a bronze rating represents failure and that puts the reputation of the UK's higher education sector unnecessarily at risk. That recommendation was based on extensive consultation. There is little evidence of similarly serious research in the reasoning given for retaining these names in the current consultation document. We support the recommendation of the independent review and would suggest that these ratings be replaced by 'Outstanding', 'Highly Commended', 'Commended' and 'Meets UK Quality Requirements'.

6. Provider and student submissions

We consider that the data-based indicators are limited in their coverage of aspects of excellence in teaching, learning and student outcomes. We note that the consultation document also explicitly acknowledges this in para

210b. This means that a detailed provider submission is essential to allow panels to form an expert opinion which is not overly affected by limitations in the available data sources.

Likewise, student submissions have the potential to provide useful evidence to the panel which may help to address limitations in the data-based indicators.

7. Indicators of student experience and outcomes

In our submission to the independent review (p.2), we argued that uncertainty needed to be properly represented in the presentation of data-based indicators. We are satisfied that the proposal in the current consultation does that. We are also pleased to see that there is no intention to flag positive or negative performance based on (inappropriate) statistical tests, a practice which we have also been critical of in previous consultations.

We find the 2.5% threshold for materiality of the difference between a provider's indicator and its benchmark to be inadequately justified. The use of the word "material" suggests that an (average) difference of 25 students per 1000 is the level at which OfS considers that experience/outcomes start to have a practically important effect. Our reading of the linked documentation suggests that this level for materiality has actually been derived to ensure that the number of providers exceeding the materiality threshold is not too high, and so really has nothing to do with materiality at all.

We note that only the value of the indicators relative to the benchmark will be presented and not the absolute values. This suggests that, at least as far as the indicator evidence is concerned, TEF is designed only to consider performance relative to the sector. If TEF judgments are to have any value in an absolute, rather than relative, sense, then it would seem necessary for information on the absolute values of the indicators to inform those judgements.

We could not find any detailed discussion of why it had been decided not to use salary measures constructed from LEO as a further progression indicator.

8. The process of expert review

Proposal 11 of the consultation sets out how the evidence used to assign a rating should be interpreted – proposing that the data-based indicators used should contribute to no more than half the evidence of excellence in either student experience or student outcomes.

As outlined in §2, the data-based indicators described are, mostly, only indirectly associated with quality of teaching, learning and student outcomes Furthermore, particularly on the student outcomes aspect, there are important dimensions for which there is no data-based indicator. Therefore, any rating process system will necessarily be driven by expert review and it is good that this is recognised.

However, the reliance on expert review does raise questions around the reproducibility of TEF results. This is an area where we have expressed concerns before in our submission to the independent review, where we argued that to ensure that results are reproducible it is vital that the methodology is fully described. The reliance on expert review to provide at least half of the evidence for a rating means that the problem of reproducibility is arguably even more magnified . Clear and consistent panel guidelines will be essential for reproducibility, and we address this issue in the next section.

9. How evidence will be assessed

Ensuring that the panel correctly, consistently and appropriately interprets the statistical evidence it is presented in the form of the data-based indicators will be critical to TEF. This will require clear, comprehensive and appropriate

panel guidelines. Unfortunately, the consultation document, specifically Annex F, does not give us confidence that these will emerge. We have three particular concerns.

First, It is stated that the indicators should "contribute no more than half of the evidence of very high quality or outstanding features, within each aspect". We agree that over-reliance on the indicators in forming panel judgments is undesirable. But it is very unclear as to how "half" should be interpreted here. How is a panel member to determine whether they are intrinsically over-weighting the evidence provided by the indicators?

Second, while we welcome the strong steer in paragraph 206 against the concept of an initial hypothesis or other formulaic judgement solely based on the indicators, we consider that Annex F does not give confidence that such approaches will not reappear "by the back door". Categorising statistical evidence as described in paragraph 15 of Annex F arguably encourages this. It also needs to be recognised that the level of evidence for a provider materially exceeding (or falling below) its threshold is a function both of the performance of the provider and its size. For two identically performing institutions of different sizes it is very much more probable that the larger provider will be considered to have provided strong evidence of that performance above (or below) threshold. This was an explicit criticism of flagging procedure in the previous TEF, pointed out in the ONS evidence considered by the independent review. Panel guidance needs to be sensitive to considering very different sizes of provider.

Our third and most serious concern, though, is one that RSS have raised on numerous occasions previously. There is insufficient attention paid to the issue of multiple comparisons. This exacerbates the concerns above around the categories in paragraph 15 of Annex F. Even if they were considered appropriate for consideration of a single indicator in isolation, they become completely inappropriate when a set of indicators is considered simultaneously.

As we understand it, even before splits are considered, there will be 15 (5 measures for 3 modes of study) indicators for student experience and 9 for student outcomes. When assessing these multiple comparisons, the evidence levels need to be recalibrated and clear and careful guidance needs to be presented. This is not mentioned in Annex F at all, and the document *Supporting information about constructing student outcome and experience indicators for use in OfS regulation: Description of statistical methods* while explicitly discussing the issue of multiple comparisons does not give confidence that this issue will be adequately addressed. Paragraph 31 of that document essentially asks users to make their own adjustments, with no further advice on how they should do so.

Clearly that is inadequate advice for a TEF panel or, we believe, for providers trying to understand their own indicators in preparing their TEF submission. When indicator data are presented on a huge number of splits, then the multiple comparisons issue is exacerbated. If examination of the split indicators is going to be anything other than a fishing expedition, clear advice is essential.

10. Transparency

We have argued previously for the TEF process to be open and transparent and the proposal in paragraph 218b of the consultation is welcome. We note however the use of the word "normally" and would expect that omission of any of items (i-iv) listed here from the published information would only ever occur in exceptional circumstances.

11. Should the scheme be called the "Teaching Excellence Framework"?

We do not agree that the wider changes which are being made to the framework in any way address the mismatch between the name and the focus of assessment. If anything the wider focus of assessment and the simpler proposed name makes this mismatch worse. We consider that a new name would indicate that the framework has, indeed, been significantly changed. The independent review proposed that TEF be renamed as "the Educational Excellence Framework", which seems to be more appropriate to us.

12. The use of statistical measures when considering a provider's performance in relation to numerical thresholds

The RSS has serious concerns about the incorrect use of statistical measures in the proposed B3 process. We have raised similar concerns with TEF but the problem is arguably even more serious in B3. The proposed B3 approach is strongly driven by identifying indicators which have strong or very strong statistical evidence (as defined in 238) of being below threshold. The problem is that the definition of strong or very strong evidence here completely ignores the issue of multiple comparisons. It is not clear to us exactly how many indicators are computed for each provider, but when all the splits are considered it seems likely to be many hundreds. When assessing these multiple comparisons involving so many indicators, the evidence levels need to be completely recalibrated and clear and careful guidance needs to be presented for providers on how this will be done. This is not mentioned in Annex F at all, and the document *Supporting information about constructing student outcome and experience indicators for use in OfS regulation: Description of statistical methods* while explicitly discussing the issue of multiple comparisons does not give confidence that this issue will be adequately addressed. Paragraph 31 of that document essentially asks users to make their own adjustments, with no further advice on how they should do so. Clearly that is inadequate advice for providers trying to understand their own indicators and gives no confidence that the B3 process will be statistically robust. It also needs to be recognised that the level of evidence for a provider falling below threshold is a function both of the performance of the provider and its size. For two identically performing institutions of different sizes it is very much more probable that the larger provider will be considered to have provided very strong evidence of that performance below threshold. This was an explicit criticism of flagging procedure in the previous TEF, pointed out in the ONS evidence considered by the independent review. This discrepancy in the way in which institutions of different sizes are treated should be addressed.