ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Submission to the Independent Review of the TEF

(Based on the TEF/RSS review listening session held on Tuesday 22nd January, part of the TEF review led by Dame Shirley Pearce)

The Royal Statistical Society (RSS) is a professional body for statisticians and data analysts, with more than 10,000 members in the UK and worldwide. As a charity, we advocate for statistics to be used effectively and honestly in the public interest.

The RSS still believes that there are many serious statistical problems with the TEF, some of which are around statistical communication. Some of these invalidate TEF.

The RSS has already responded in detail to two previous TEF consultations run by the Department for Education. See:

https://www.rss.org.uk/Images/PDF/influencing-change/2018/RSS-Evidence-Dept-Education-Teaching-Excellence-Framework_final-21May-2018.pdf

and

https://www.rss.org.uk/Images/PDF/influencing-change/2016/RSS-response-to-BIS-Technical-Consultation-on-Teaching-Excellence-Framework-year-2.pdf

Our main points are summarised below. The RSS has raised many of these before, often in our consultation responses and some have not been addressed in successive TEF incarnations.

We repeat our conclusion from the 2018 Subject-level consultation, which remains our position, although TEF now seems to be the responsibility of the Office for Students (OfS):

"*The Royal Statistical Society (RSS) was alarmed by the serious and numerous flaws in the last Teaching Excellence Framework (TEF) consultation process, conducted in 2016. Our concerns appeared not to be adequately addressed by the Department for Education (DfE). Indeed, the DfE's latest TEF consultation exercise, which will shortly close, suggests that few statistical lessons have been learned from 2016's [or 2018's] experience. As we argue, below, there is a real risk that the latest consultation's statistically inadequate approach will lead to distorted results, misleading rankings and a system which lacks validity and is unnecessarily vulnerable to being 'gamed'.*"

When we refer to TEF documentation below, we are largely working from two recent documents published on the OfS website on 22nd October 2018 (OfS 2018.44). They are the *TEF Subject-level pilot guide* ("the pilot guide") and the *TEF Guide to subject-level pilot data* ("the pilot data guide")

A. *Uncertainty handling*

(i) There is occasional mention of uncertainty in the TEF specification, usually to do with specific and limited tests and procedures. However, there is no overall coherent and consistent treatment of uncertainty throughout the whole of TEF. This applies to the statistical methodology and the subsequent human committee that transforms TEF metrics and flags into TEF awards (Gold, Silver, Bronze). Quantifying and communicating uncertainty in the statistical process is a minimum requirement and well-established in the statistical literature. It is more difficult to ascertain this in the human panel evaluation part of the process: expensive, but not impossible.

Ultimately, the RSS judges it to be wrong to present a provider/subject as Gold/Silver/Bronze without communication of the level of uncertainty. The current TEF presentation of provider/subjects as Gold, Silver, Bronze conveys a robustness that is illusory. A prospective student might choose a TEF Silver subject at one provider instead of a TEF Bronze at another institution. If they had been told that, statistically, the awards are indistinguishable, then their choice might have been different and, in that sense, TEF is misleading.

The fact that uncertainty is, apparently, not properly and holistically assessed and not communicated appears to transgress Q3.3 of the UK Statistics Authority's Code of Practice for Statistics [UKSA-CP] ("The extent and nature of any uncertainty in the estimates should be clearly explained").

The uncertainty is likely to be higher for subject-level assessment than for provider-level assessment, due to the smaller samples involved and that, for provider assessment, the diverse human panels might have stronger knowledge about an entire institution rather than for specific subjects at that institution.

(ii) Accurate and coherent uncertainty assessment is also vital to understand the value and cost-effectiveness of the TEF. If it turns out that the uncertainty swamps the mean-level award (Gold, Silver, Bronze), then this calls into question whether it is even worth continuing with the TEF. Even if the signal-to-noise ratio is acceptable, and some statistical meaning can be attached to the mean-level award, TEF still might not be worth the cost. A robust cost-benefit analysis should be carried out and subject to the usual independent scrutiny, e.g. the National Audit Office.

(iii) UKSA-CP V5.5 concerns the burden on the sector ("The burden imposed should be proportionate to the benefits arising from the use of the statistics"). Since we do not know whether TEF is statistically effective (from (ii)), or have a detailed and robust cost-benefit analysis, it is our view that UKSA-CP V5.5 has not been satisfactorily adhered to, particularly at the subject level.

B. *Comparability*

Is a TEF Gold at one university the same as TEF Gold at any other university?

The answer has to be no. One reason is that the TEF Award is predicated on the use of benchmarking (or benchmark groups) and different missions of groups at institutions. Benchmarking is used because TEF realises not all universities have the same mission or teach the same kind of materials and it is an attempt to control for differences in demographics. Statistically, TEF Gold at one institution can not necessarily be compared with TEF Gold awarded to another. This is potentially deceptive and misleading for stakeholders, particularly students.

If benchmarking were achieved using pre-selected groups of similar institutions (e.g. Russell Group), then one could imagine a TEF award being qualified by the benchmark group. For example, TEF Gold (Group 1), TEF Gold (Group 2), etc. To use an Olympic metaphor, people understand that a Gold Medal in shooting is very different to a Gold Medal in volleyball but understand that Gold is indicative of the highest level of achievement.

An analysis might be performed that sought to discover whether the benchmarking factors might be used to identify reasonable university groupings and, if they did, it might be possible to qualify TEF award with a benchmark group. Overall, though, we suspect this is unlikely.

The use of the same TEF award, and the same TEF logo, for all types of university seems highly misleading. The literature and communication around TEF should make it clear that TEF awards are not comparable across the board. This goes for both provider- and subject-level processes and would enable TEF to adhere to the UKSA-CP Q3.3 "The quality of the statistics and data, including their … **comparability**, …, should be monitored and reported regularly"

C. *Benchmarking*

We are extremely worried about the entire benchmarking concept and implementation. It is at the heart of TEF and has an inordinately large influence on the final TEF outcomes.

(i) The RSS has referred to benchmarking in the past as a 'poor person's propensity analysis'. By this we mean that TEF uses some observed covariates to 'stratify' the results (so Cambridge gets 'compared' at some level to Oxford, but not to universities that are quite different). The *pilot data guide*, we believe, admits this "Where differences exist … they may be due to some other characteristic, which is not included in the weighting" (para 72, p20). This is precisely our point: differences in TEF metric scores might be due to *unobserved* characteristics unrelated to teaching quality. So, attributing the differences to teaching quality is unscientific and wrong.

In some statistical situations, randomisation is used to protect against the influence of unobserved covariates. It would not be ethical to use randomisation here, but the unobserved covariate problem still exists and should not be ignored. At the very least, it should be strongly communicated to stakeholders, especially prospective students, that the TEF is observational in nature and that TEF differences are likely not solely due to

teaching quality differences. Proper communication would enable adherence to UKSA-CP Q3.1: "The limitations of the statistics and data should be considered in relation to different uses, and **clearly explained alongside** the statistics."

(ii) Para 76 of the *pilot data guide*: "The benchmarking methodology seeks to ensure that student and course characteristics that have the largest effect on what we are measuring". TEF benchmarking does not include important characteristics such as amount of course content, diversity (in its broadest sense) or difficulty/challenge of material. Surely, this has an enormous effect on what is measured?

This seems wrong in itself. We are concerned that omissions of this sort will lead to game playing by institutions. One might improve NSS scores, for example, by 'dumbing down' the syllabus and there is strong anecdotal evidence that this is already happening in the sector. Ultimately, such game playing will not be good for students (on the 'market', against competitor students internationally), nor for the providers' reputation and ultimately deleterious to the economic well-being of the UK.

(Indeed, OfS already has evidence of unexplained grade inflation:

https://www.officeforstudents.org.uk/publications/analysis-of-degree-classifications-over-time-changes-in-graduate-attainment/

which might be evidence of 'dumbing down' or related behaviours. How much of this is stimulated by exercises such as TEF or NSS?)

(v) Para 80 of the *pilot data guide*: why is 2% materially different? And why should it be 2% for all metrics?

(vi) Para 82 of the *pilot data guide*: on the UKPI method for testing significance of metric differences.

The exact details of the process are hard to come by. We felt that they were not properly explained by reference to the HESA website (footnote 36). As part of proper reproducibility, we think DfE/OfS should include a section, or separate document, that fully describes the methodology (see point D, next).

We are also seriously concerned about the use of the benchmarking method for the TEF and whether the underlying assumptions made are valid. We are in the process of investigating this with actual TEF data, but have experienced difficulties on obtaining details on the data and the process.

In particular (a) it is not clear that the data forming Z-scores are, or should be, normally distributed or even symmetric. Some data stem from categorical variables and previous DfE TEF consultation documents rightly indicated various clustering behaviours. That something unusual is going on seems to be borne out by observing a curiously large number of Z-scores attributed to some indicators. (b) The benchmark itself is a random quantity. Is this taken into account? (c) Is the joint distribution between benchmark and indicator of a provider/subject taken into account and modelled? These and other questions could be answered if DfE/OfS adopted a fully transparent approach.

A major problem with the TEF use of the UKPI methodology is that the Z-score is calculated, and assessment is performed, individually for each indicator, for each subject, for each provider for each metric. Overall, there are huge numbers of these, as can be seen in the TEF data spreadsheets, distributed by DfE. The Z-scores are assessed as being statistically significant, at some prescribed significance level. However, these multiple assessments seem to be performed individually and separately using the same critical value as for a single test. No allowance has been made for size control of this (massive) multiple hypothesis test, which will result in hugely more significant results and many more flagged indicators than is warranted.

At Dame Shirley's listening session, the RSS enquired of the DfE/OfS representatives whether multiple testing without adequate size control was occurring and the answer seemed to be yes.

Since this seems to be the case, then this lack of overall size control is a serious statistical mistake and means that many (previous) TEF flags should not have been so flagged. This appears to transgress UKSA-CP 2.1: "Methods and processes should be based on national or international good practice, scientific principles, or established professional consensus."

Wikipedia has entries on this 'look-elsewhere' effect and some examples of where this misuse has caused erroneous scientific conclusions. The 'Data dredging' page should also be consulted, which refers to this kind of incorrect analysis as "the misuse of data analysis to find patterns in the data that can be presented as statistically significant when in fact there is no real underlying effect".

We believe that this is happening here: far too many flags are being raised, erroneously alerting the downstream human TEF panels to effects that are just not there. Our conclusion is that the previous TEF awards are not valid.

D. *Transparency and Reproducibility*

Many in the community, including the RSS, would like to see the processes and data underlying the TEF put fully into the public domain. This would enable several assumptions made about the data and process, by OfS/DfE, to be tested. It might also permit processes such as the TEF to be improved, if that is possible. As mentioned in C(vi) above there are many questions relating to the benchmark process itself and also the data used to calculate them. Helpfully, the OfS does provide some information on its website (Home>Advice and guidance>Teaching>TEF data>Get the data) in the form of some spreadsheets that provide some information. For example, we can see many of the Z-scores, but, in general, it has been extremely difficult and time consuming to delve deeper. It should not be like this.

At a minimum, we would expect the entire TEF data process pipeline to be published, including as much data that can be released ethically. We have reports of people (in and outside the RSS) trying to understand the TEF data release, but find the accompanying instructions impenetrable. There is a lack of transparency, which is fuelling a perception of lack of integrity.

At the moment, we believe that the current data/process release is not adhering to several items in UKSA-CP V2 Accessibility and UKSA-CP T6.1: "Relevant nationally- and internationally-endorsed guidelines should be considered as appropriate". For example, the UKRI Common principles on data policy and/or the Transparency and Openness Promotion guidelines (https://cos.io/our-services/top-guidelines/)

E. *Small sample sizes, missing and non-reportable data and categorical data*

There is often not a lot of detail on how small sample sizes are going to be handled or where metrics are missing or non-reportable. It does get a mention but, in these situations, a lot depends, presumably, on the judgement of the analysts who are producing the TEF outputs (flags, indicators). The problem is that these questions arise in many (thousands) of cases and nearly all require detailed human statistical judgement, not automatic processing. However, we just don't know.

For example, consult Table B1 on page 80 of *the pilot guide*. In that table missing data on specific questions seem to be completely ignored in the calculations appearing in the final column (% agree). The 'strongly agree' and 'agree' categories used in computing the outcome are merged. This raises the question of what is the point of differentiating between 'agree' and 'strongly agree' in the first place? There is a body of literature in survey methodology that suggests that respondents tend to pick middle categories when asked questions on a Likert scale. Is this something to be concerned about?

Another example is how the issue of non-reportable metrics are going to be handled, not only in themselves, but comparing subjects/providers with different levels of missingness, and again communicating that through to the end result.

F. *Does the Teaching Excellence Framework really assess teaching excellence?*

Fundamentally, do the metrics input to TEF measure quality of teaching? Do the provider submissions measure teaching quality? We are sceptical. There may be some distant indirect association, but what robust research been carried out to assess this? Alternatives might be to rename TEF (to remove 'teaching excellence'), or actually carry out some evaluation of teaching quality (which would be expensive).

We do think it is useful for students to see the metrics that underpin TEF, relating to their potential course choice. The Unistats website already does this and seems to be useful and well-used by potential students. The RSS could imagine an upgraded Unistats site containing well-chosen and well-communicated metrics being valuable for prospective students and other stakeholders. Caveats with individual statistics can be indicated on the site near to those statistics (cf UKSA-CP Q3.1 as mentioned above). Such an approach has the advantage of directly feeding each student's utility function, rather than externally imposed implicit utility imposed by TEF. It might be argued that the TEF's philosophy that distils diverse institutions into three categories, underestimates the intellectual ability of prospective students and other stakeholders.

G. *Gaming and Goodhart's Law*

What explicit, practical, steps are being taken to detect and prevent gaming of TEF?

H. *Lack of Multivariate Assessment*

The multivariate nature of the metrics, interaction with panel assessment and teaching quality, does not seem to be captured by TEF.

I. *Lack of consideration of time effects*

TEF also does not appear to capture the time series nature of teaching quality. We have made this point previously in our consultation responses. What is the evidence to say that a teaching quality mark now will result in a student getting a good experience in several years' time? This is related to point A(ii) on cost-benefit analysis.