Evidence does not support the rationale of the TEF

Graham Gibbs

Abstract

The Teaching Excellence Framework (TEF) has evolved since it was first announced, and HEFCE guidance to institutions on its implementation reveals a number of significant concessions to evidence, common sense and fairness. Institutions may well implement useful teaching improvement mechanisms in response, as they have always done, regardless of the nature of external quality assurance demands. However, the rationale of the TEF remains – and it is deeply flawed. It is the rationale that this paper focuses on. It is argued here that its interpretation of evidence about educational quality, employability and value for money ratings, used to justify a TEF, are irrational and are not supported by evidence. Making fine judgements about institutional rankings (and hence fee levels) on the basis of metrics is likely to be thwarted by the very small differences in scores between institutions. Some of its proposed metrics are invalid. Its belief in the ability of a small panel of experts to make sound quality judgments is not well founded, given the poor record of past attempts to make such judgements about teaching quality in higher education. The higher education market is very complex and perhaps only a minority of institutions will be able to benefit in the way the TEF intends. The TEF seems unlikely to be perceived, by most, as rewarding.

The Teaching Excellence Framework's underlying assumptions

However unfit for purpose past teaching quality regimes have been, they have often resulted in institutions' putting more effort into improving teaching than previously, because the risks of not doing so have been perceived to be significant. Most institutions have markedly improved their National Student Survey (NSS) scores since the NSS and metric-based league tables were introduced, under a regime that has focused on quality assurance rather than on quality and under which fees have not been linked to quality. Institutions seem likely to take the TEF extremely seriously, whatever they think of it. The TEF is built on a number of explicit assumptions stated in the Green and then White papers. It is argued here that these assumptions are unfounded. If teaching quality does improve, and it might, it will not be because Government policy is soundly based.

Current teaching metrics do not indicate that there is a substantial teaching quality problem that needs an urgent solution

The government argues that the TEF is necessary because teaching quality is unacceptably low. However, the NSS, that provides one of the only ways currently to monitor quality over time, reveals a completely different picture. Levels of satisfaction and judgements about teaching are high and have gone up every year but one since it was introduced. Those scores that were initially lower (such as for assessment and feedback) have shown the largest improvements. The rate of improvement shows little sign of slowing even though there is a ceiling effect – the scores are often so high there is little room for further improvement. It is true that a few institutions (especially elite research universities such as Imperial College and the LSE) have performed less well recently. It is also true that students are very generous judges of teaching: about three quarters of all teachers are usually considered 'above average'. The picture the NSS provides is probably too rosy. However, the overall trend is inescapable. A more credible interpretation of the available quality data is that existing teaching metrics, however flawed, have been surprisingly successful in levering institutional efforts to improve teaching quality, particularly outside the research elite, even in the absence of variable fees. It might be the case that collating the data differently (for example, not bundling together the top two ratings on rating scales, which tends to exaggerate how good things are) or adding new and more valid quality data (such as concerning students' level of engagement) would provide even more effective leverage to institutional efforts to improve. But there is nothing in the existing evidence that points to the pressing need for varied fees as a lever on the grounds that otherwise institutions will do nothing to improve.

Poor 'value for money' is not caused by poor quality

It is argued that alarmingly low 'value for money' ratings justify a strong emphasis on improving teaching quality. But satisfaction and teaching quality ratings are very much higher than value for money ratings! Low value for money ratings are to do with high cost, not low quality. Whatever the level of quality, the cost of higher education is perceived as too high because the much cited 'graduate premium' (the additional income graduates can expect simply as a result of being a graduate) is unrealistic. It is based on historical data when there were fewer graduates, the economy was expanding and wages were higher in real terms. It is not at all clear that the current economy needs the current number of graduates each year and this is reflected in the proportion of graduates who, at least initially, undertake non-graduate-level jobs

with low wages. The cause of perceived low value for money is that students, quite realistically, are worried that they may not be able to recover their very substantial investment. The experience of the USA, with many graduates never repaying debts caused by ever-higher college tuition fees, provides a perfect example of how it can all go wrong. Higher fees reflect higher reputations, but, in the USA, reputation predicts almost nothing about teaching quality or student learning gains or the extent of use of educational practices known to improve student learning gains. High fees have become a proxy for high quality – but they are a thoroughly misleading proxy.

If 'value for money' is low, it makes no sense to put fees up

Faced with ample evidence of perceived poor value for money, the rational thing to do (if you are incapable of improving the economy and the employment market) is to lower the investment students need to make – to reduce fees. Instead the government say they will improve value for money by increasing fees, at least for most. This is *Alice in Wonderland* logic.

The higher education market does not work perfectly or uniformly

The TEF naïvely assumes a perfect and uniform market. It assumes that all institutions would seek to raise fees, and would raise them if they were allowed to, and that they would automatically benefit as a result; by doing so, they would increase both their attractiveness to consumers and their income. This ignores the reality that many institutions operate in local or not very flexible markets, in which prospective students may have little choice about where to study or much flexibility over how much they can afford to pay. There are already examples of institutions which have increased fees, to take advantage of excellent NSS scores, only to find that they cannot fill their places and have had to put their fees down again. Some institutions, even with comparatively low fee levels and/or perfectly respectable teaching quality metrics, are currently not filling their places. Those institutions that recruit nationally and internationally may benefit from an increase in perceived reputation that comes with higher fees and be able to exploit their market; many others cannot do so, however good they are at teaching. There are assumptions about the market that make some sense for the elite but not for others.

Increased income from raised fees may have little impact on teaching quality

There is an assumption that increased income from increased fees would be spent on further improving teaching. The overall evidence about the relationship between income and teaching quality suggests that the link is weak, at best. In the USA, tuition costs have doubled, and

doubled again, with no improvement in class sizes or other valid teaching quality indicators, and current tuition costs are unconnected with educational quality. In the UK, the comparatively richer Russell Group Universities actually have larger cohorts and larger classes than do poorer 'teaching-intensive' universities. Russell Group students experience a smaller proportion of academic-led small group teaching, because graduate teaching assistants are so often used to save academics' time for research activities. Cohort size, class size and the proportion of teaching undertaken by people other than academics are all good negative predictors of student learning. The research elite do not, in the main, spend their money on teaching students if they can help it and there seems little prospect of a change in their policy, wherever the money may come from.

Distinguishing appropriate fee levels for institutions is unreliable, and in the homogeneous middle range, impossible

There is an assumption that it is possible, safely and fairly, to make fine-grained distinctions between institutions, so that a range of fees can be fixed in precise relationship to a range of teaching quality. Three significant problems prevent this assumption from being remotely reasonable, the first two being associated with the two forms of evidence that will determine decisions: qualitative judgements by panels and quantitative metrics.

In the TEF's first stage, there are due to be qualitative judgements (basically a 'yes' or a 'no'), made by some kind of expert (or inexpert) panel, about whether institutions deserve to be allowed to put their fees up. Those who are as long in the tooth as I am will remember Teaching Quality Assessment. Every subject in every institution in England was allocated a score out of 24 as a result of qualitative judgements made by a large panel of 'subject experts'. The process involved visits, observation of teaching and meetings, often with teachers and students, and the collation and examination of truly vast piles of documentation. It took six years to implement. Despite the enormous cost in time and effort, the extensive evidence base, the visits, the training of assessors and so on, the outcomes were highly unreliable. Some subjects allocated much higher average scores than others, with no discernible justification. Later scores were higher than early scores. Most scores were so high as to be indistinguishable for the vast majority of institutions. But, more worryingly, there were substantial systematic biases. There was a strong positive correlation between research strengths and TQA scores, despite its being known that research strengths do not predict teaching quality. What is more, larger departments and institutions gained higher scores than did small ones, despite the fact that size was known

to be a negative predictor of educational quality. It seems that assessment panels were dazzled by reputation and were incapable of making reliable judgements about teaching. The TEF's qualitative judgements are intended to be made extraordinarily quickly, by small panels without the benefit of visits, observation, meetings or even detailed evidence about educational quality. Instead, they will largely be looking at a very short text prepared by institutions themselves. The chance of their making sound and precise judgements seems negligible; the chance of their being dazzled by reputation seems somewhat higher.

The second problem facing the TEF relates to attempts to make distinctions between institutions about what level of fees they will be allowed to charge, on the basis of teaching metrics. I can still visualise the graphs I was shown twenty-five years ago, when the Course Experience Questionnaire (CEQ) was first used in Australia to provide public comparative quality data about every subject in every university. The CEQ is (or at least was, until the Australian Government turned it into a 'happiness' questionnaire) a valid instrument for judging educational quality. It produces scores on a range of credible variables and is adequately reliable and valid. The graphs I saw took one scale on the questionnaire at a time (such as 'deep approach' – the extent to which students attempted to make sense of subject matter rather than only to memorise it) and ranked every department in the country in a particular subject. What was immediately clear was that a couple of departments were measurably worse than the rest at the bottom and a couple were measurably better at the top; everyone else was pretty much indistinguishable in the middle. This was true for every scale on the questionnaire, for every subject. Statistically, this is an inevitable consequence of the variable being measured being more or less normally distributed and with a small standard deviation - a phenomenon apparent for virtually every variable about quality one can think of when comparing institutions. With NSS scores, the same is true. If you look at national rankings for 'satisfaction', you find the vast majority of institutions in an indistinguishable middle, with adjacent institutions having almost identical scores, and even blocks of ten institutions not differing significantly from adjacent blocks of ten. No less than forty-three institutions shared NSS satisfaction scores of 85-87% in 2016. You can tell an institution ranked 120 from an institution ranked 20, but not one ranked 50 from one ranked 60. The differences are so small and so volatile from one year to the next, that overall rankings can change markedly, year on year, without any change in the underlying phenomenon. Such variations are picked up by 'The Times' and trumpeted in such emotive headlines as "*University X crashes down quality league",* when in fact the change in score has been random and statistically insignificant. It is rarely possible to distinguish one institution from

the next in a reliable and safe way using such metrics because the differences are, in most cases, simply too small. Yet that is exactly what the TEF has to do – say that one institution deserves to charge higher fees whilst the next one down the rankings does not, even though they are statistically utterly indistinguishable. Adding scores together from a bunch of varied, and often invalid, metrics actually makes this problem worse and produces a grey muddle. The HECE guidelines now suggest that no more than 20% of institutions might be indentified at the top and bottom of rankings and distinguished from the middle-ranked institutions. But even that is bound to create unfairness for the institutions just below the boundaries that will be, in any statistical sense, indistinguishable from those just above the boundaries.

Institutional average scores on teaching metrics usually hide wide departmental differences

The third problem facing the TEF in making the required fine-grained distinctions is that it intends to rank and distinguish institutions. Institutions are made up of departments (or subjects) that very often differ widely from each other in terms of a whole range of metrics. These internal differences can be so large that an institution may have the top-ranked department in the country in one subject and the bottom-ranked department in a different subject. These departmental scores are then averaged and the institution as a whole might end up looking average (as most in fact do). This averaging of varied departments helps to produce the problem of lack of distinction between institutions highlighted above. Students need to know about the subject they are interested in and to be able to compare that subject across institutions. The current TEF mechanism will not allow them to do this - it could even trick them into paying a higher fee to study in a lower quality department. If students are interested in their likely employability, the problem is even more acute, as national differences between subjects are gross, and institutional employability averages are, at least in part, a consequence of their subject mix. If an institution taught just Nursing and Cultural Studies, then it might look average for employability, but comparatively bad for a student wishing to study Nursing and surprisingly good for a student wishing to study Cultural Studies. This problem would be partly solved if the TEF operated at the level of subjects (or departments) rather than institutions, which is a development being considered for the future.

But even then, there would be significant difficulties in identifying what a ‘subject’ is. I once helped a Sports Science department collect a good deal of data about students’ experience of assessment, using the Assessment Experience Questionnaire (AEQ). There were seven degree

programmes within 'Sports Science' and, in terms of students' experience, they differed from one another to a considerable extent, ranging from rather good to pretty awful. As there were no NSS categories to differentiate between these degree programmes, for NSS data collection and reporting purposes the seven were simply aggregated into a single undifferentiated muddle. Standard NSS subject categories might work for traditional academic departments with one degree programme and large cohort sizes, but they may be less than helpful as a means of distinguishing the more unconventional and varied subject groupings usually found in modern teaching institutions. Again, this suits traditional research universities best.

Employability has little to do with teaching quality

There is a misinformed and confused conflation of employability with quality. Quality, according to the TEF, is apparently all about employability. Students don't think so. Those responding to a HEPI survey asking them what best indicated the quality of a course had some perhaps surprisingly conventional ideas about teachers and teaching; employability came nearly bottom in their reckoning in terms of telling them anything useful about quality.

If the government had bothered to look at national rankings of universities' teaching and employability performance, it would have discovered that its assumption is complete nonsense. The table below ranks institutions according to 2016 NSS scores.

Table 1: Institutional teaching and employability rankings

Rank	**NSS** 2016	**Graduate employability** Times, 2016	**NSS rank** 2016
1	Buckingham	Cambridge	20
2	University of Law	Oxford	20
3	St Marys College Belfast	LSE	155
4	Courtauld Institute of Art	Manchester	87
5	Keele	Imperial College	116
6	St Andrews	Kings College	129
7	Bishop Grossteste	Edinburgh	145
8	Harper Adams	University College	102
9	Liverpool Hope	London Business School	155
10	Aberystwyth	Bristol	76

It will be noticed that most of the top ten institutions are neither prestigious nor research giants. The second set of rankings is from 'The Times' 2016 data collection about graduate employability. There is no overlap at all with the top ten for NSS satisfaction. The right-hand

column lists the NSS rankings for the top ten institutions for employability. With the exception of Oxford and Cambridge, they are considered by students to be amongst the worst in the country. Imperial College led the clamour for higher fees; it is currently ranked 116th for student satisfaction and dropping like a stone, but its reputation guarantees effortlessly-high graduate employability metrics.

It took about ten minutes to compile this table from data easily available on the internet.

Employability is largely a product of reputation which follows research performance, overall income and visibility. Graduate employability has almost nothing to do with teaching quality and most institutions are not in a position to do much about the employability of their students, which is largely determined by employers' notions about reputation and the employment market - often the local employment market. The government could do something about that, but not universities.

It is also the case that size helps visibility and reputation, and hence employability, but hinders teaching quality. It is rare in research literature about good teaching departments to discover one that is even medium-sized, let alone large. Top research universities are mainly large and tend to keep students' choice of courses down, so creating large cohorts and large classes in order to reduce teaching loads. The consequences are there for all to see.

The TEF's proposed teaching metrics have limited validity

The TEF rests on teaching metrics' being valid. If they are not, in the sense that they do not predict student learning, then orienting institutions to improving them may distract institutions from actual efforts to improve student learning. The government is fully aware of the contents of 'Dimensions of Quality' (Gibbs, 2010) and its identification of which metrics are valid and which are not, and so the proposals in the original Green Paper about the metrics the TEF would use were guaranteed to dismay. By the time details of the implementation of the TEF were made public, the situation had improved. Nevertheless, 'satisfaction' is not a valid measure of learning gains or of teaching quality. Outcome measures (including retention and employability) are significantly determined by student selectivity, and so indicate reputation rather than teaching quality, and reputation does not predict learning gains or the extent of use of pedagogic practices that lead to learning gains. The introduction of benchmarks that take the nature of student intake into account will help here, and 'The Times' modelling of institutional rankings based on benchmarked TEF metrics, using 2015 data, produced somewhat inverted rankings

compared with the newspaper rankings we are used to seeing (that have been created by using almost entirely invalid metrics). This cannot have been what was originally intended. It is possible that the TEF's benchmarked metrics, even if some of them are invalid, will create quite a shock to the system. Increasing the role played by valid measures, such as of student engagement, will help in the future and it is to be hoped that there will continue to be pragmatic changes in implementation in the pursuit of validity and fairness. The first attempt to produce rankings and associated varied fee levels is unlikely to get it right and decisions about institutions' futures based on the current form of implementation are likely to be dangerously unsound. It would be prudent to wait until some of the problems identified above have been tackled more satisfactorily and to treat the rankings of the first year or two as a wake-up call. It is not as if students are impatiently pushing the government hard to increase fees.

The TEF is unlikely to be perceived by most as a reward

The government argues that, just as strong research performance is rewarded by the REF, strong teaching performance should be rewarded by the TEF. But the majority of institutions have seen their research income decline dramatically over several rounds of research selectivity. The REF and its predecessors were designed explicitly to allocate research funding to fewer institutions (and fewer researchers) and to take funding away altogether from most. Careers, working lives and institutional reputations have been blighted by the REF. For most, it has been experienced as a punishment. Similarly, the TEF is likely to be perceived as offering brickbats and an uncertain future to perhaps thirty institutions, and as damning by faint praise perhaps 100 more. Only those institutions that are allowed to charge top whack, and the sub-set of these for whom this is actually useful and welcome, are likely to feel rewarded: big sticks and small carrots, again.

Conclusion

A national policy with this degree of leverage over institutional behaviour risks causing damage if the assumptions on which is built are wrong and the measures it uses are invalid. Institutions may feel obliged to play the system and try to improve their metrics even if they do not believe in them and even if this has no useful impact on student learning. But perhaps institutions will become more sophisticated about using appropriate metrics in sensible ways. The demands of the TEF for evidence of 'impact' are already stimulating fresh thinking. If that prompts new evidence-based approaches to enhancement, then the TEF might even improve students'

learning gains despite its rationale and design. As students will be paying even more for their education, let us hope so.

Reference list

Gibbs, G. (2010) *Dimensions of Quality*. York: Higher Education Academy. Available at: file:///C:/Users/Home/Desktop/General/HEA%20Quality/Dimensions%20of%20Quality%20Final%20Report.pdf. (Accessed: 19 January 2017).