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

Annex C: Subject panel reports

This is an independent report completed in autumn 2019 following the conclusion of the pilot.

Publication date 21 January 2021

Contents

Introduction 1

Arts and Humanities subject panel report 2

Business and Law, and Education and Social Care subject panel report 15

Medical Sciences, and Nursing and Allied Health subject panel report 27

Natural Sciences, and Engineering and Technology subject panel report 40

Social Sciences, and Natural and Built Environment subject panel report 49

Subject panel reports for TEF subject-level pilot 2018-19

Introduction

The assessment of subject provisions in the second Teaching Excellence and Student Outcomes Framework (TEF) subject-level pilot (2018-19) was carried out by 10 panels of subject experts. Each panel consisted of student representatives, academics and employer representatives, including members of professional, statutory and regulatory bodies (PSRB). The 10 subject panels were paired for the exercise to create five combined subject panels, as shown in the table below. Each combined panel was co-chaired by two academics and supported by two students as deputy chairs.

Subject panel structure		
Arts	**+**	**Humanities**
Business and Law	**+**	**Education and Social Care**
Natural and Built Environment	**+**	**Social Sciences**
Medical Sciences	**+**	**Nursing and Allied Health subjects**
Natural Sciences	**+**	**Engineering and Technology**

Full details of the assessment process are outlined in the TEF subject-level pilot 2018-19 guide[1] published in October 2018. Comprehensive evaluation of the second pilot is reported in 'TEF: Findings from the subject-level pilot 2018-19', to which this report is an annex. Details of the panel membership are published on the Office for Students (OfS) website.[2]

[1] Available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-subject-level-pilot-guide/.

[2] Available at: www.officeforstudents.org.uk/advice-and-guidance/teaching/future-of-the-tef/subject-level-pilots/.

Arts and Humanities subject panel report

Executive summary

Introduction

The Arts and Humanities Panel (AH panel) comprised 16 academics, nine student members and four further members representing employers. The panel met on 25 to 27 March, 29 April to 1 May, and 30 May 2019 to produce subject-level ratings under the piloted model of the Teaching Excellence and Student Outcomes Framework (TEF). The panel was co-chaired by Professor Susan Orr, Dean of Learning and Teaching Enhancement and Professor of Creative Practice Pedagogy at the University of the Arts London and Professor April McMahon, Duty Vice-Chancellor (Education) at the University of Kent. With James Perkins, Former Vice-President (Education) of City University London Students' Union and Peter Cowen, Vice-President of The Open University Students' Association, as Deputy Chairs. The panel assessed 155 submissions across eight subject areas:

- Combined and area studies (2 submissions)
- Creative arts and design (32 submissions)
- English studies (25 submissions)
- History and archaeology (20 submissions)
- Language and areas studies (12 submissions)
- Media, journalism and communications (22 submissions)
- Performing arts (31 submissions)
- Philosophy and religious studies (11 submissions)

Key findings of the panel

- There was a clear consensus from the panel that the focus of TEF on enhancement was positive.
- The panel noted that there were challenges to using TEF as a student information tool. The panel acknowledged that the 'no rating' award may need to be given in certain cases for reasons other than the panel deemed there to be insufficient evidence to make a 'best-fit' judgement.
- Student panel members played a key role in the delivery of subject-level TEF by working in partnership with academic panels members.
- The revised criteria for subject-level and provider-level assessment were welcomed and helped the panel reach robust judgements.
- The panel struggled to have confidence in judgements made for cases at the lower threshold for inclusion in subject-level TEF (student populations of approximately 20).

- The panel had questions over the value of Longitudinal Education Outcomes (LEO) data for measuring student outcomes particularly in relation to graduates working in creative industries.
- The panel welcomed the opportunity to rate each aspect of quality, and to provide a Statement of Findings (SOF) to providers, though the utility of more detailed statements needs to be balanced against scalability.

Evidence, including metrics and evidence in submission

This section focuses on the assessment criteria and their application. The differentiated criteria for provider and subject assessment were welcomed by the panel and the new criteria (TQ1 and TQ5) helped the panel focus on student experience, partnership and engagement.

The panel noted that student engagement was understood in two ways: firstly as student engagement in the subject teaching and learning contexts, and secondly as student engagement in the preparation and co-authoring of the subject-level TEF submission. There were different views among panel members about how much could be inferred from one to the other. Arguably, the panel's view of what constitutes student voice was not completely settled.

Some subject panel members felt that it was important to see the student declaration when agreeing a rating, and that it would have been helpful to have these declarations available for all subjects. Others argued that the student declarations can, in some circumstances, serve as proxies for student engagement as a whole. Some students on the panel wanted to see evidence in all cases that students had been involved in writing the subject submission and were very keen that submissions would be understandable and accessible to a prospective student audience.

Rating descriptors and boundaries

The panel's discussions around rating descriptors focused on two boundary points. The first was the Silver/Gold boundary. The panel deliberated the phrase 'sector leading' (as a criterion for the Gold rating) because the phase was not part of the rating descriptors. Reference to this phrase sometimes resulted in a submission near the borderline coming down to a Silver rating.

The other boundary that was the source of much discussion was at the lower end of Bronze. Here discussion focused on what constituted the lower threshold for a Bronze rating.

The section of the TEF criteria that refers to 'positive outcomes for all' (SO1) was also the focus of debate. The panel discussed this criterion in relation to providers that recruit students who live locally after graduation in areas of deprivation and low wages. The 'outcomes for all' discussions focused on graduate outcomes because attainment data is not available at subject level.

One panel member suggested that it would be useful to have opportunities to have space for shared discussion of the descriptors for Gold, Silver and Bronze ratings at the start of the process. This would have supported the development of reflective, shared understandings of what each descriptor might mean and what each descriptor might look like for a range of subjects. This could be a more explicit part of the calibration process.

Metric workbook and contextual data

It was interesting to recognise the ways that providers and panel members were responding to the extended metric workbook. There was a concern that the sheer wealth of data led to panel members and providers being selective in the areas they focus on. However, as a whole, the contextual data was used in a similar way to last year. Panel members incorporated the additional data into their overarching judgements and evaluation.

Maps were rarely referred to and when they were used it was in relation to providers in areas of low employment. Maps were usually considered in cases where it was necessary to help secure and confirm a rating that was close to borderline.

Metrics

Whilst the National Student Survey (NSS) weightings were widely accepted several panel members questioned the double weighting of the continuation metric, stating that the rationale for this double weighting was unclear.

The Longitudinal Education Outcomes (LEO) data was the most contested part of the metrics workbook. Some panel members felt very uncomfortable equating the quality or value of a degree with salary outcomes and they did not feel confident that benchmarking was satisfactory in this area. It is also noted that this data does not reference the employment and enterprise outcomes of international students who make up a large percentage of the student body in some universities. It could be concluded that LEO data in TEF produces a disincentive for institutions to recruit international students. Challenges pertaining to LEO data in the context of the creative industries are explored later in this report.

The panel questioned the relevance of subject submissions that referred to the providers' current employability support as a means to contextualise data concerning students who graduated nearly ten years ago. The panel noted that none of the submissions referred to historic career support and advice.

Missing data

The panel's approach to missing data changed slightly this year from the first year of the pilot. This year it was compulsory for providers to make a submission; consequently, the panel had a strong view to judge the subject on the evidence present and there was less emphasis on the view that 'absence of evidence is not evidence of absence'. The majority view was that providers should, as far as possible, receive a rating if they submitted metrics that met the OfS threshold and offer the two-page provider submission and five-page subject submission.

Attainment data

It was noted that on the main panel the use of the attainment differential data was rarely referred to in relation to 'outcomes for all' and this was a concern given the urgency of the sector's need to address attainment differentials across its split metrics.

Z-scores

Colleagues valued having access to the Z-scores of both 1.65 and 1.96. This information was used as part of holistic judgement but providers rarely make use of Z-scores in their submissions.

Subject expertise

The panel recognised that the construction of a larger, multi-disciplinary panel from the previously separate Arts and Humanities subject panels was a temporary measure for this year, as part of a test of scalability for a full-scale subject-level exercise. However, the panel welcomed this way of working, particularly because many had expertise across more than one Arts and Humanities subject, and indeed across the previous panel divisions. As a result of this wide-ranging expertise, disciplinary expertise appeared to be referenced less in the judgement process in Arts and Humanities panel than other panels; it was certainly not an area of contention.

The experience of the Arts and Humanities panel is that members contributed fully and equally, and felt confident in doing so, regardless of their role on the panel and their professional background. The process was a respectful and good-humoured one and there was only one case where there was a clear difference of view between academic and student members. This case was debated fully and with a focus on its more general implications of a divide in opinion between panel members as well as debating the specifics of the case. It was fundamental to the panel that the students were heard and persuasive in determining panel ratings. The chairs and deputy chairs were sensitive to all student feedback throughout the panel's processes to ensure that any concerns expressed in separate student panel member meetings were addressed in the Arts and Humanities panel. The students were fully integrated and respected members of the panel.

The panel did have concerns about diversity in terms of ethnicity and sector representation for academic and student panel members. Specifically, further education colleges and alternative providers were less well represented than the panel would have liked for a substantive exercise. The panel expressed concern that a number of student representatives, who have a number of years of experience as TEF panel members, are now some distance from being students. It was agreed that this could represent a challenge in terms of losing substantial expertise all at once, and the panel was keen that consideration be given to ways of scaling up student involvement and training student panel members. This would not involve just the mechanics and ethos of TEF assessment but crucially 'rules of engagement' in large meetings and strategies for intervening confidently and ensuring voices are heard. This is all the more important if more students per se, rather than student representatives such as current or recent sabbatical officers, are to be brought into the process, which the panel would support strongly.

In terms of student voice in provider submissions, the panel was disappointed that the additional guidance given to providers had not been reflected in an improvement overall in commentary about student voice. The student voice was hard to hear in many submissions and was absent altogether from others. There was considerable discussion within the panel about different means of representing and including student voice, and agreement that further guidance and good practice examples may have helped providers. Likewise, the assessment phase would have benefited from consideration of how the student declarations should be used at subject-panel level, when it is currently primarily a matter for the main panel. This could have been in the form of a mandatory section on student engagement in subject submissions.

Most subjects in the scope of the Arts and Humanities panel are not externally accredited, and therefore the panel made minimal use of professional, statutory and regulatory body (PSRB) factors and had no members specifically representing PSRBs. The panel did, however, agree that the employer representatives brought substantial added value to discussions and additional rigour to the panel's judgements. In particular, there was recognition of the important role of employer representatives in articulating and testing subject initiatives against industry norms, to distinguish expected good practice against innovation. Employers regularly participated as additional readers, even when they had not initially read a case, and their contributions in assisting us to tease out some highly complex interactions between local employability landscapes and subject submissions were greatly appreciated. The panel felt it was important that employer involvement be a formal part of the assessment process, so employers were clear about their input being sought and recognised, and we avoided duplication. Not having a specific caseload meant employers were free to choose their preferred cases, but this meant that some submissions were not considered by an employer representative. The panel felt the assessment process would have benefit from more structured employer input, for example using a set of questions devised between the OfS and the employer representative which could be used to ensure we are making the most of the experience our employers bring to the process.

While the involvement of employer representatives was universally felt to be extremely valuable, it was harder to evaluate the contributions of widening participation (WP) and interdisciplinary liaison members of the panel. Not all cases were seen through a WP or interdisciplinary lens in terms of specialist or liaison reading, though it is fair to say that equality and inclusivity issues were a topic of regular debate and particular awareness for the panel. However, it is worth noting that the WP and interdisciplinary liaison leads gave helpfully steer to the panel in several cases. The issue with assessing interdisciplinarity provision, in particular, was how should the panel assess information once it had been highlighted, and here the panel would have appreciated further guidance.

There was considerable discussion, and also a sense of lack of clarity, about the relationship between TEF and access and participation plans. Here the panel sensed a potential for overlap in terms of governance and reporting. Interdisciplinarity continues to be an issue for Humanities in particular, with many students having a joint honours experience, and yet are referred to in subject submissions relatively rarely. Given the importance of interdisciplinarity in Arts and Humanities it was felt that the interdisciplinary liaison role could be further developed with clearer OfS guidance regarding how to employ the additional data sets that the liaison role has access to. Further guidance is recommended for providers and subjects, to ensure they demonstrate the experience of students working across disciplines is being considered and attended to.

Inclusive TEF processes

Assessment approaches deployed as part of the TEF have to enable all panel members to contribute equally, particularly due to the large volume of cases which were, and would need to be in a scaled exercise, assessed in a fixed window. The panel ensured that the needs of all panel members were met and that expectations about reading requirements were set out in a timely way to support inclusion.

Submissions

There continued to be concern from panel members who noted that some providers were misinterpreting what is required of them in relation to engagement with their metrics workbooks. The panel noted that some providers simply summarised the data or challenged the data or offered their own version of the metrics that did not match the metrics workbook. In common with provider-level TEF and earlier stages of the pilot it was observed that some providers preferred to state what was on offer to students rather than providing data in relation to take up and impact on learning. Panel members who were part of the pilot last year did not see improvement in the general quality of submissions this year.

The Arts and Humanities panel was concerned about the language used in some submissions when students with protected characteristics or broader WP backgrounds were referred to. In some cases, negative or patronising language was in evidence and sometimes submissions tended to refer to certain groups of students to explain or mitigate metrics in a way that hints at a deficit view of the student. The panel would have appreciated a mechanism to feed this concern back to providers.

Student information and subject-level TEF

The general consensus in the panel is that TEF is less likely to fulfil a student information agenda due to the bluntness of the three point rating approach, the benchmarked metrics workbook which makes cross-institutional comparisons problematic, and the challenge of ensuring that submissions cover criteria in a way that is accessible and jargon free for prospective students.

In addition, changes in course provision, provider mergers and provider restructuring make it hard for students to make inferences about the future based on current TEF ratings. The panel was, however, absolutely committed to the utility and importance of TEF for the enhancement agenda.

The assessment process

The panel valued the balance between metrics and submissions in reaching an overall, holistic judgement and was committed to constant interrogation of the relationship between these different components of evidence. The panel was conscious to avoid 'metrics capture', which might lead to an over-reliance on one evidence source. Comparison of the panel's initial ratings at the beginning of the assessment against the eventual holistic rating, and cross-panel comparisons, were helpful in maintaining awareness of avoiding an overreliance on metrics.

On the whole, the panel considered that the risk of over-reliance on metrics was appropriately and successfully mitigated. It was observed that there were differences in proportions of final ratings between the two assessment batches carried out by the panel. However, the panel was satisfied that this reflected genuine differences in quality of provision between the cases in each batch, albeit recognising limitations in confidence due to the relatively small sample size. The panel remained conscious of the relatively low percentage of Gold ratings awarded and we return to this issue below.

On the relatively small number of occasions where the panel awarded a ‘no rating’, or where this was discussed as a serious option, the key factors were:

- missing data, often reflecting very small populations, and often combined with a relatively 'thin' and underspecified subject submission;
- a lack of information or even comment about substantial cohorts of students who were in scope for the subject but did not feature in the submission;
- a small number of cases where the panel agreed a rating was not appropriate.

When there was limited or missing data the panel considered that the introduction of a minimum population of 20 students did allow for testing of minima in terms of our capacity to produce robust ratings. However, 20 is a very small population and created real challenges in efforts to award a rating wherever possible. The panel agreed that a minimum population of 40 would have been preferable and would have supported the robustness of subject-level TEF. The panel did, however, recognise that this would create a challenge for inclusivity, especially for further education colleges and alternative providers, where assessment may only be possible at provider level.

In terms of submissions not commenting on students in scope for the subject but not featuring in the submission, the panel took the clear and consistent view that it is exceptionally important for subject submissions to show knowledge and understanding of their students. Where significant proportions were not mentioned (whether because they were undertaking a specific programme of study, or were joint honours students, or part of a cohort at a partner provider), then the panel struggled to agree a rating.

There was some concern in the first batch of assessment that, due to the considerable workload, there could be a tendency for the panel to settle for a majority decision rather than pushing for consensus. This might mean the panel agreed a rating too soon, rather than keeping cases on the borderline between two ratings open through the assessment process. The panel took this seriously and sought to investigate and mitigate for this by returning to a number of 'borderline' cases, for further consideration and discussion. While there were no lasting concerns over individual cases, it is important to stress that parity of rating was regularly part of the subject panel's agenda.

The panel discussed at length the relationship between provider-level and subject-level ratings. It was recognised that a provider with an overall rating that was not shared by subjects might encounter adverse comment. However, the panel considered that an important step had been taken in last year's subject level pilot to remove the automatic read-across between the proportion of the subject in relation to the provider’s overall provision. It was noted that the processes for awarding subject-level and provider-level ratings are separate and different. Any impression of 'correcting' subject-level ratings because they don't 'line up' with the rating awarded at provider-level risked devaluing the contribution of subject specialism. However, the panel did note with interest that the involvement of a subject specialist assessment by no means correlated with the award of a higher rating for the subject.

The panel welcomed changes to the Common Aggregation Hierarchy at level 2 (CAH2) for the second pilot which led to fewer unclear and multidisciplinary categories. It was still clearly possible to reflect the experience of students taking more than one subject, and the better submissions did this successfully. There was some discussion about cases of 'leakage', when the panel is considering several subjects from the same provider, and finds it cannot un-know information from one subject when rating another. The panel found the inclusion of a list of programmes of study particularly helpful; this contributed to our understanding of the experience these students were having in terms of pedagogy and subject expectations (as some of the subject classifications

themselves in Arts and Humanities are very broad), and also allowed us to better understand cases of new programmes being introduced, or others being taught out. The panel welcomed the list of programmes contributing to each subject and found it a useful addition to the process.

The panel took a lively interest in options for moderation, and made full use of additional readers looking at cases that were contentious during the panel discussions. While there was different practice across subject panels, the Arts and Humanities panel took the clear view that it was important for the additional readers to be aware of the locus of disagreement which we were asking them to resolve. When additional readers were sought, in cases where the original three assessors and other contributors had not been able to reach an agreed view, the panel provided a briefing to the additional reader with an opportunity for questions about the nature of the disagreement. This had the advantage of allowing additional readers to be assigned in terms of expertise and interests relevant to the cases. It enabled the panel to ask for more than one additional reader when it was felt that a particular perspective would be useful. The Arts and Humanities panel therefore operated a system, in academic and quality assurance terms, of applying moderation rather than a full double-blind second-marking process. While there was still robust discussion of some of these borderline cases the panel tended to trust the judgement of the additional readers and quickly came to see this as a very important aspect of the process.

The panel initially adopted the approach of allowing additional readers an hour to cover their new cases in the afternoon during the assessment meeting ahead of further panel discussion about the case later that day. The panel moved to a revised system during the second set of assessment meetings; cases with additional reading were discussed on the following morning, giving overnight for reading. While obviously this made for more out-of-hours working, the panel accepted that this was entirely reasonable for the short periods of meetings and regarded this as a small price to pay for the greater inclusivity arising from giving additional readers more flexibility. Quality and confidence of fourth reader contributions noticeably improved in the second set of assessment meetings. This allowed for all panel members to feel equally confident to contribute to additional reading.

In terms of the general dynamics of meetings, the panel felt that further consideration needed to be given to the workings of the three initial assessors for each case. In the subject panels, as opposed to the main panel, the three initial assessors did not actually meet and were not given access to one another's rankings or any comments in advance of panel discussion. The assessors therefore came into the discussion without knowledge of whether the case was 'clear-cut' and required little discussion, or whether it was highly contentious and would need a more extensive scene-setting presentation from the lead reviewer. In consequence, the panel felt time was wasted on covering entirely straightforward cases in detail, when we could have taken a more risk-based approach more confidently and allocated time better for the challenging and borderline cases. We are aware that there is a challenge of scalability in factoring meetings in advance, though the process may have benefited from virtual discussion between the initial assessors, or through sharing of data even if discussion were not possible. There was a sense in this year's subject pilot that we were not quite making full use of the panel’s time in general, with a lot of time for most panel members spent listening rather than actively participating, especially in stages where half of the panel is involved. This is arguably not the best use of panel members' time or expertise, though of course we recognise the need for the whole panel to be signed up to eventual ratings decisions. Changes to streamline the panel assessment process may have allowed for greater scalability and allow panel members a fuller sense of agency in the process.

Ratings and statements of findings

The Arts and Humanities panel, while welcoming a number of new members, was relatively constant in its membership between the first and second subject pilots. While most members were therefore accustomed to working with the three Gold, Silver and Bronze ratings and descriptors, there continued to be disquiet through the process about the consequences of this system. There was great concern about the 'cliff edge' effects of a very small number of ratings, where so much rests on which side of the great divide a provider or subject will end up on. At the same time, the variation between the top and bottom exemplars of a category is vast. Panel members had one robust discussion about whether quality thresholds could confidently be said to have been reached.

In consequence, the panel warmly welcomed the capacity to keep subjects on a Bronze/Silver or Silver/Gold borderline open until later into the assessment process, allowing for a more detailed follow-up discussion. In the second batch of assessments, ratings for each of the three aspects of quality were allowed to stay on these borderlines throughout and statements of findings (SOFs) use a five-point scale for the three aspects of quality. The panel would have appreciated the same options for the overall subject rating, with five ratings having substantially more support than three. While ratings descriptors would have to be revised for a five-point scale, the proximity to boundaries would be far less catastrophic. In addition, the burden of scalability would be considerably reduced by not having to map from a five-point to a three-point scale at subject level.

The no rating option was welcome, though the Arts and Humanities panel consciously sought to minimise its use and to award a rating wherever possible. This was in part due to a concern that 'no rating' might have negative connotations, and however hard we might work to explain that this is typically a reflection of a lack of data, it might well be interpreted as a case assessed as below Bronze. The panel felt that this problem would continue, and particularly for small populations below 40.

The panel was pleased to be able to recognise some exceptional practice but felt that the proportion of Gold ratings was not quite where we might have expected it to be. There was some disagreement and debate, ratings descriptors notwithstanding, over quite what 'outstanding' means, and whether outstanding provision has to be absolutely sector-leading. Likewise, the panel was somewhat conflicted over how literally 'for all students' should be taken. These debates, while recurrent, were positive in character, and the panel was confident that robust and defensible ratings had been achieved. The relatively low proportion of Gold ratings may be an artefact of the particular sample of providers and subjects in this pilot. However, an eye needs to be kept on this in any scaled-up subject TEF. The panel noted that the basket of metrics made achieving Gold in the early stages of assessment exceptionally difficult compared to Bronze or Silver/Bronze, to the extent that the panel tended to burst into spontaneous applause when a subject 'made it' on the core metrics. Given our deep conviction of the high international quality of the UK higher education system, we feel it would not be unreasonable to equalise the chances of starting as Gold or Bronze on the metrics. Panels are perfectly capable of demoting as well as promoting cases on a holistic basis, so we feel the risk is small.

The option of awarding a rating for each of the three component aspects of quality was welcomed warmly by the panel, who sensed that this provided greater robustness and confidence in our ratings. The only difficulty we faced in respect of the three aspects of quality lay in the configuration of the learning environment (LE) aspect, where learning resources and continuation did not seem

to fit coherently together. If this aspect of quality included academic community (with a focus on sense of belonging), along with academic support, which together enhance the prospects of continuation, this might make for a more coherent section.

The panel was inconsistent in reaching the overall ratings for the subject and the aspects of quality. In some cases, the panel struggled initially to reach an overall rating but were able to work our way to an agreed position via the ratings for the three aspects. In other cases, the overall rating seemed clear, but the panel were able to express nuances through differentiating at aspect level. The panel benefited considerably from this flexibility, and while we are aware that some panels took a particular view of the necessary direction of travel between aspects and subject, this panel also considered whether rating the aspects of quality cost us time, and concluded that it did not, overall. While the process was somewhat longer for some straightforward cases, because four scores had to be agreed rather than one, it definitely saved us time in many more complex cases.

The matter of ratings for aspects of quality is also relevant to the choice between the two types of SOFs. The panel was greatly appreciative of the return to provision of SOFs, which were not part of the first subject-level pilot. If the key function of TEF is the enhancement agenda, then helping providers and subjects to understand the back-story to their rating is vital. And this is not antithetical or oppositional to an argument about benefit for students, either. While we do not believe TEF will be much used by prospective students, who already have a wealth of information to call on in making their choice of higher education provider, a successful enhancement within a provider or subject will benefit everyone – staff, current students, and students yet to come. So enabling enhancement is crucial, and that means giving providers and subjects good information to help them understand their rating, and to move forward and improve. However, while the panel welcomed the opportunity to include some limited additional text in the second batch of cases, albeit under closely controlled circumstances, we would have preferred this to be clearly communicated back to providers/subjects, not only to the main panel.

This is, of course, the classic burden versus value argument. The panel appreciated that producing standardised SOFs was easy and swift, with a straightforward choice from a bank of statements, yet the panel preferred the narrative-style SOF, where more detailed and nuanced feedback is possible. As chairs and deputies, we also see the risk of free text SOFs, which for our panel were highly variable in detail, length and clarity. This makes them clearly more contestable. The panel was consoled by the inclusion of the ratings per aspect of quality, feeling that these added value for subjects and providers in identifying the loci of (relative) strengths and weaknesses; but there was still a clear wish to add some extra text at least in exceptional cases. It seems to us quite feasible to converge on a model for SOFs which is intermediate between the first and second batch, with ratings per aspect, perhaps a bank of additional statements, but a de-risked and less burdensome process compared with the first batch. The panel felt that the quality of SOFs was potentially compromised by the difficulty of simultaneously capturing the panel discussion about a subject while completing the assessment process.

The panel also noted that the relationship between ratings for aspects of quality and for the subject overall are not in a linear or entirely predictable relationship. It is possible to have the same three scores for per aspect ratings which map to different subject ratings. The panel was generally relaxed about this, given the availability of borderlines and a five-point scale for the aspects of quality, which then had to map onto a three-step scale at subject level.

The panel did note some variations in performance between subjects, such that some areas (Philosophy and Religious Studies; Languages and Area Studies) received no Gold ratings. We asked whether there might be a correlation with proportion of joint honours, with many languages departments, for example, having moved away from a single honours model. However, we do recognise that the populations, in terms of the number of pilot subjects and providers, is likely too small to settle this question.

Sector-wide enhancement

The time given to the deliberations and decision making for each provider did not allow for careful collation and analysis of best practice. The process would have benefited from an OfS officer or panel member who was responsible for recording examples of best practice to be compiled and shared. Many examples of best practice can only be understood in relation to the provider type and context, which would be complex to extract and write up.

Subject-specific observations and additional considerations

It was notable this year that there was a greater focus across the submissions on decolonising the curriculum initiatives. A small number of providers referred to their work addressing their attainment differentials, which are indicative of the growing work in this field. The panel would like to have seen addressing attainment gaps as an important focus for TEF. There was much less enthusiasm for the inclusion of 'grade inflation' as a metric, as this was felt to be causally ambiguous.

Supporting enhancement

The aspect ratings will help providers identify areas of strength and areas for enhancement. Panellists who were part of the pilot stressed that the conversations taking place in their institutions as a result of involvement were feeding directly into enhancement activity in a positive way. This points to an overall positive effect in the equalisation of education and research in the sector and in specific institutions.

The panel noted the enhancement potential for TEF in helping identify and potentially address such important topics as decolonising the curriculum and addressing attainment gaps.

It was noted that the Arts and Humanities panel seems to respond more positively to innovation rather than to the ordinary, possibly applying values beyond the 'Frameworks' criteria. There was concern from the panel that enhancements that build from years of development could be penalised because they are not strictly innovative.

Potential impact

Panel members from the creative arts and performance subjects continue to be concerned about the unintended consequences of subject-level TEF in relation to creative arts education. The government has a focus on STEM subjects – Science, Technology Engineering and Mathematics. Several commentators have argued for the inclusion of arts so that the acronym becomes STEAM. This is to recognise the centrality of arts in education and industry. These concerns were discussed in last year's report and are summarised here:

- Changed approaches to measuring school attainment in the English Baccalaureate are already leading to reductions in resourcing, staffing and opportunities for pupils to study drama, art and design. The reduction in numbers of pupils studying in these areas is already leading to many further and higher education providers experiencing dips in recruitment.

- Creative arts and design courses need specialist space and are staff and resource intensive. Creative arts education is particularly vulnerable in providers managing budget constraints resulting from under recruitment.

These contexts could lead to a position where a Bronze rating for Arts in a multi-subject provider could be used as a further rationale to close courses. We note that this works both ways and Gold subject ratings will help secure funding but we are keen to communicate the ways that lower ratings in this subject will add to an already unstable context in relation to pipeline and resourcing. We note that similar arguments can be made for resource-intensive subjects in the Humanities which are currently challenged in terms of student recruitment, notably languages.

There are also wider structural employment contexts that need to be considered in relation to the use of LEO data as a measure for the creative industries; some of these concerns are also applicable to a number of humanities subjects.

Creative arts and design are a very large subject area, studied by 9.6% of higher education students in the UK. It is unusual as a mainstream discipline in being delivered mostly by small, specialist institutions. Its graduates almost all work in the creative industries, now worth more than £100bn per annum and growing at twice the rate of the economy as a whole. Despite the surge in graduate-level jobs, the structure of the labour market in the creative industries holds the average wage of creative workers below the repayment level for student fees.

The last major longitudinal study of creative higher education and careers was 'Creative Graduates, Creative Futures' (2010) by the Institute for Employment Studies.[3] It reviewed the early career patterns of more than 3,500 graduates across 26 institutions in practice-based art, design, crafts and media subjects. The employment characteristics of the creative industries in particular include a high concentration of creative graduates and micro-enterprises combined with low average income for creative workers. However, graduate earnings remain an incomplete measure of individual motivation for career and study choice. The evidence suggests those who select creative subjects are less motivated by income than other factors. For example, one university with a large group of creative arts students worked with Youthsight to explore motivations for study in more detail. The findings suggested that potential earnings do matter to creative students, but as a secondary factor. What matters more to them is acquiring the knowledge and skills to realise their chosen career path and being able to find employment in their sector of choice. Career happiness has long been important to students choosing creative art and design. It was found to be a key motivation for study in the study 'Creative Graduates, Creative Futures'.

Research into the creative industries exposes the ways that regional and sectoral differences matter, and will influence individuals' career choices. Research by Creative and Cultural Skills showed that mean pay for creative occupations is higher than mean pay for the total UK workforce:[4]

[3] Available from: www.employment-studies.co.uk/resource/creative-graduates-creative-futures.

[4] Available from: https://ccskills.org.uk/supporters/advice-research/article/workforce-analysis-2018.

‘Some of the surprising information to come out of the research surrounds pay. Wages are generally thought to be low in the cultural sector compared to other industries, but this seems to be dependent on where you’re working. The average hourly wage for the UK’s whole working population is £14.77 an hour, while for the cultural sector is actually £16.29 an hour. In London, you’ll earn on average £19.48 an hour if you work in our sector compared to £18.63 for the rest of the working population there. Whereas in the North East you’ll earn on average £10.77 in our sector, whilst the rest of the working population there earns £12.96.’

Acknowledgments

The Arts and Humanities panel is confident that the subject-level TEF framework was applied appropriately and that judgements made for each subject in the pilot were carefully considered and robust.

The panel chairs want to take this opportunity to thank the deputy chairs, James Perkins and Peter Cowan, for their substantial contribution to subject-level TEF, and indeed to acknowledge one another's contribution and companionship.

In addition, we note the generous contribution made by Derek Hamilton and Ruby Gatehouse.

Authors

Professor Susan Orr

Professor April McMahon

Business and Law, and Education and Social Care subject panel report

Executive summary

Introduction

The Business and Law, and Education and Social Care panel (BLESC panel) comprised 14 academics, eight student members and four further members representing employers. The panel met on 26 and 27 March, 30 April and 1 May, and 30 May 2019 to produce subject-level ratings under the piloted model of the Teaching Excellence and Student Outcomes Framework (TEF). The panel was co-chaired by Professor Julia Clarke, Pro Vice-Chancellor (Business and Law), Manchester Metropolitan University and Professor Dilly Fung, Pro Director (Education), London School of Economics and Political Science. With Michael Olatokun, student representative at BPP University and former Student Union Community Officer, University of Nottingham, and Rebecca Maxwell Stuart, former Vice-President (Education) of University of Stirling Students' Union, as deputy chairs. The panel assessed 109 submissions across four subject areas:

- Business and management (36 submissions)
- Education and teaching (25 submissions)
- Law (22 submissions)
- Health and social care (26 submissions)

This section provides a review of the pilot process and outcomes for the BLESC panel. The section on student views has been written by the deputy chairs and provides additional feedback solely from the student members of the panel.

Overall, we consider the panel made robust assessments at both the aspect and overall level. The ratings generated through holistic judgements accurately reflected, on the basis of the processes we were required to follow, the provision assessed.

The panel worked very effectively both as a whole and in smaller sub-groups. The range of discipline expertise and experience of different provider types was valuable.

The panel identified some modifications to the process that would have benefited their assessments, most particularly that:

a. more assessment could have focused on subject 'outliers' and

b. the Gold, Silver and Bronze ratings could have been more usefully replaced with Good, Excellent and Outstanding.

Evidence, including metrics and evidence in submission

Metrics and evidence

The introduction of a minimum cohort requirement in this pilot was judged to be a positive. The new guidance on using 'no rating' in cases where there was not enough evidence to make a fair

and rational decision was welcomed. For most cases the panel had a full or near full set of data to consider when making judgements. The panel, however, found that the minimum threshold would have been more effective if revised upwards. This would have avoided efforts being put into cases by both providers and reviewers where no rating was the only outcome that could be reached with sufficient assurance.

The panel endorsed the continued emphasis on experience of, and outcomes for, **all** students. Again, the minimum cohort requirement reduced the incidences where split metrics were not shown.

The pre-calculation of the step 1a hypothesis was helpful to panel members. Nevertheless, metrics and contextual information have become complex and very time-consuming to consider, especially where a dual hypothesis is required. This raises a danger of 'metric fatigue'. Reviewers may become overly focused on a subset of the quantitative data provided, avoiding that which they find less easy to understand and to interpret. In particular, the majority of panel members did not find the maps helpful. We also found no purpose to the inclusion of greyed out values given that panel members were not meant to refer to them.

It was noted that if the existing basket of metrics and contextual information continues into the post-pilot stage, new reviewers will have a very steep learning curve.

The panel appreciated that the initial hypothesis should not be interpreted as indicative of the final rating. Nevertheless, it considered that the patterns of flag combinations that determined the initial hypothesis were problematic. These were too skewed towards Bronze or Bronze/Silver ratings and away from Gold and Gold/Silver ratings. This means that reviewers have to consciously guard against a deficit mind-set in coming to their holistic judgement.

As in the previous pilot, some panel members considered that there was an over-emphasis on employment metrics which accounted for 3 out of 7.5 (40 per cent) of the metrics. This was particularly problematic when using the second method trialled, whereby reviewers were required to make judgements against the different aspects. Panel members felt that student learning outcomes and learning gain were too narrowly defined through the emphasis on employment outcomes. There was also a strong view that continuation could usefully be viewed in connection with the student outcomes (SO) and learning gain (LG) aspects, rather than in the learning environment (LE) aspect.

Similarly, the 'student voice' criterion would seem to fit better under LE rather than teaching quality (TQ). Such a re-positioning could help to promote the importance of the student voice, and of working in partnership with students, in enhancing their education and their wider learning environment.

The Longitudinal Educational Outcomes (LEO) metrics were deemed to be problematic because of the time lag between the institution's historic practices and the outcomes data. The lack of regional benchmarking for the 'above median earnings threshold' metric was also thought to be flawed.

Evidence in submission

As in the previous pilot, the written submissions were of mixed quality. The better written submissions took an evaluative approach. Those that were weaker were often overly descriptive and made too much use of selective quotations. Stronger submissions demonstrated the provider's

awareness of the need to address particular metrics and an excellent understanding of their student population. They also used other data sources effectively to provide evidence where metrics were missing or where innovations were not yet reflected in the metrics.

Some of the variability in the quality of the submissions appears, as in the previous pilot, to be due to differences in institutional capabilities and resources. Panel members were mindful that some providers have the benefit of cadres with significant experience and understanding of the 'rules of the game' of higher education review. Others do not, or they may be used to different systems of regulatory oversight. The panel considered that more dissemination of good practice in writing subject submissions ahead of the exercise would have helped level the playing field.

The panel discussed the use of direct quotes from NSS or external examiners and were concerned that this may be contrary to General Data Protection Regulation (GDPR) requirements, as this data had been originally provided for a different purpose. As NSS comments are not supposed to be used for marketing, the question was raised as to whether they should be admissible in TEF submissions where these will become public. The panel requested guidance from the OfS on this as part of its final evaluation and feedback.

Any form of external accountability based on annually reported performance indicators creates a danger of short-termism. Those being held to account may be rewarded for temporary fixes over permanent cures. One aspect that the panel did debate was how much credit should be given to positive actions that are clearly in train but have yet to demonstrate impact. The panel also considered it to be important that providers should be able to feel confident to be frank about interventions that had not worked. Stronger submissions clearly demonstrated that providers' attempts at enhancement, successful or otherwise, were rooted in an understanding of their students' lived experiences. These concerns might be addressed through the ratings descriptors making explicit reference to 'strategic commitment to enhancement'. This might also provide a means for drawing out the student voice more explicitly.

Weaker submissions are still not fully capturing the contribution of students to enhancement. Panel members noted greater use of free text comments from surveys and feedback mechanisms used to monitor and evaluate courses than in last year's process. Selective quoting of this sort rarely provides robust evidence of the student voice in the enhancement of learning and teaching. As such, it tended to be of little value to the panel in assessing the commitment to student engagement in enhancement.

Some submissions did reference where students had been involved in the development of written submissions. Others included sections written by students. Signs of an authentic strategic involvement of students strengthened judgements, but its absence was not penalised. The panel was mindful that for some types of institution it is challenging to capture the student voice.

The student declaration had been added as an additional component for the 2018-19 pilot year, but subject panels did not use them in their assessments. It was hypothesised that this may have resulted in subject-panel members not being able to see the full picture with respect to student voice, and that it could have been better if student declarations were read by the initial trios of assessors.

The panel's chairs and deputy chairs who had access to the student declarations as part of the main panel assessment process felt that the student declarations raised a number of issues:

- Although the guidance relating to student involvement had been strengthened, there remains a tension between the value of data sharing and warnings about data protection.
- Students may wish to have more freedom to express themselves in a format that suits them.
- Students may not feel free to give negative feedback when working collaboratively on a submission. However, it was acknowledged that students did find opportunities for more negative comments in some of the declarations received this year.
- It was noted that a separate student submission which was very critical of an institution might not be in the best interests of the current students.
- It could be difficult to reconcile a negative student declaration with positive metrics – a challenge previously also experienced by those involved with Quality Assurance Association (QAA) review.

With regard to the usefulness of different types of data, individual members of the panel had some suggestions for additional metrics: a measurement of overall student satisfaction; an indicator of grade inflation at subject level; and the proportion of staff holding teaching qualifications.

The assessment process

General

As in the previous pilot, the panel was generally confident that it was making effective use of the complete dataset in order to arrive at holistic judgements. The combined larger panel worked effectively both as a large group and when broken into smaller sub-groups.

Trios

Initial consideration of each submission was by a sub-group of three panel members (referred to as ‘trios’). The diversity of the panel membership meant that we were able to assign cases to trios of panel members with good representation, containing at least one subject specialist. Panel members felt more secure in their own subject specialism, but different voices were important in coming to secure judgements.

Whilst the changing membership of trios had positives, it eliminated any possibility of their having a preparatory meeting. Creating fixed trios, as in the provider-level TEF exercise, would enable prior discussion, either virtually or in a physical meeting. This would be particularly valuable when individuals are unavoidably absent from the panel sessions.

Given the need to manage conflicts of interest, it was helpful to designate a trio member to take notes to inform the writing of the statement of finding (SOF). Going forward, it is to be hoped that the new systems invested in by the OfS will enable trio members to have access to notes taken by TEF officers, chairs and deputy chairs in compliance with confidentiality requirements.

Individual assessments

The revised individual assessment template worked well, allowing assessors to work step by step through the assessment process while coming to a holistic judgement. Assessments were

uploaded in a timely manner and were available for chairs and deputy chairs to see in advance of the meetings.

Panel meetings

There were some good examples of full discussions and consensus decisions being reached in stage two of the assessment process. There was sufficient time for detailed discussion of cases where there was not an initial consensus.

While it was understandable for trios not to view each other's assessments during the individual assessment stage, panel members' feedback favoured having some form of discussion prior to the panel meeting. This would have enabled a more focused discussion on the specific points of contention where individual ratings were quite different from those reached in the final rating.

There were difficulties noted in assigning a lead reviewer simultaneously to present a judgement and to record additional comments from fellow panel members that could be relevant to the SOF. In future, it would be best to separate out the function of note taking and leading the discussion, to ensure that all views are evenly captured.

Sometimes an individual panel member's stage 1 rating may have been quite different from the final stage 3 rating. This may not have been fully resolved in all cases, and preparatory meetings of the trios would have provided an opportunity to present a case in which all perspectives are clearly conveyed.

It is possible that the panel reached an 'averaging out' consensus with some of the judgements, where differences could have been resolved more confidently if more time had been available to consider the nuances of the different arguments.

However, fourth (and sometimes fifth) readers were valuable when consensus could not be reached within the trios. Fourth readers were given guidance on areas to review in the assessment rather than acting as a blind second marker. This was appreciated, as often fourth readers were asked to focus on a specific area of the submission that the trio had found challenging in reaching consensus.

Second batch assessments

Determining a rating for each aspect of quality, as well as an overall rating, led to more discussion but also more divergence. This divergence seemed to be largely on the aspect ratings, with consensus more easily reached on overall ratings.

The per aspect ratings facilitated the capture of views on specific elements, meaning that they were not lost within an overall rating that might tend to the mean of the judgements on the three aspects.

Some members of the panel expressed concerns about the relationship between the 'per aspect' ratings and overall 'best fit'. It will be important if aspect ratings are published that readers understand that they are not to be reconciled with the overall judgement through a simple additive process.

The use of borderline ratings for individual aspects of quality received a mixed response. Some assessors felt that borderline ratings led to a tendency to assume that the movement could only be upwards or downwards to the next category, while some appreciated the flexibility.

Employer expertise

As in the previous pilot, employer contributions were critically important to the assessment process and the significant commitment of employers' time and expertise was widely praised. It was noted that if the pilot is to be rolled out, there will be a need to consider carefully the commitment that employers are asked to make and how their time is best used.

The responsibilities of employers – for example, how many and which cases they should read – need to be clear from the very beginning. The employers on the BLESC panel were keen to take on a significant volume of reading and this certainly meant that they were engaged throughout the panel's discussions. This may, however, create resource issues in subject-level TEF as it is rolled out.

The time commitment for training was challenging for employers, and again time commitments need to be carefully considered.

One possibility for reducing the time commitment for employers would be for them to attend stage three of the assessment process remotely.

Widening participation expertise

The panel generally welcomed the introduction of the widening participation (WP) liaison role. There is a need, however, to better define responsibilities over and above those of other panel members. The panel has a collective responsibility for WP to be foregrounded in its considerations, yet the specialist role is a significant commitment on top of the individual's case load.

Attainment and progression outcomes for different groups of students were considered by all panel members to be a very important aspect of the process.

There are characteristics that are not identified through the split metrics – for example, students who are transgender, carers or commuting. Providers could be encouraged to comment on these issues in their written submissions.

There was discussion about the possible tension in awarding Gold or Silver ratings to providers with negative flags on split metrics. Those involved in the TEF process clearly understand the concept of 'best-fit' judgements. If, however, subject-level TEF is to help all applicants make informed decisions then different outcomes for different groups need to be clearly highlighted.

Stronger submissions wove inclusivity and WP throughout the written statements. They demonstrated an understanding of their students and of how they were tackling gaps at the micro level.

Access and participation plans (APPs) have been changed since the last pilot. The panel agreed that consideration of the relationship between subject-level TEF and narratives and targets included in the APPs could have enhanced its decision-making.

Interdisciplinary expertise

Again, the panel welcomed the designation of the new role of interdisciplinary liaisons, who were also full members of the panel. They were able to make helpful contributions to the discussions. There is a concern, however, that interdisciplinarity may need to be more carefully conceptualised. The difference between interdisciplinarity and multidisciplinarity was not always clear. This meant that interdisciplinarity did not always feature as much as it perhaps should have in submissions or discussions. One suggestion is that providers might be asked to include a paragraph about interdisciplinarity in their submissions.

Student contribution

The BLESC student deputy chairs felt that they were well supported and that they felt empowered to lead the work of the panel throughout the process. This was echoed by student members of the panel, who reported an ethos of respect and parity between panel members at meetings.

Full participation of panel members was heavily reliant on their resources; members required laptops in order to conduct the assessment of cases. If subject-level TEF is to be scaled up, panel members would benefit from the opportunity to rent computers. Further, students would benefit from being informed of these IT requirements ahead of time.

Further considerations that could facilitate the participation of students include:

- Presentation of information in a more accessible manner, particularly for panel members with specific needs
- Checking in with students to ensure that they are able to meet deadlines and attend meetings
- Understanding that not all students will have access to an office suitable for confidential work
- Ensuring that the marketing material for the recruitment exercise emphasises the opportunities and training the process provides, as many potential panel members may erroneously feel that they are unsuitable for the role.

Concerns were raised that TEF ratings are largely derived from outcomes for domestic and EU students, despite the fact that a wider international prospective student population are likely to use the TEF when making choices about UK higher education. The use of NSS data relating to international students was viewed as positive in this regard.

Quality and robustness of the assessment process

For the split metrics, it was suggested that there should be additional guidance about ‘acceptable’ levels of poor outcomes for certain categories of students. Careful consideration should be given to the level of tolerance allowed at each rating.

It was felt by some panel members that the double weighting ascribed to continuation is quite hard to overturn. Others argued that there should be emphasis on this metric, as it related to the ‘outcomes for all’ criterion.

The panel discussed whether we should try to find an alternative source of evidence for each of the aspects, to create more balance. For example, teaching quality relies on metrics derived from the NSS. The panel suggested data on teaching qualifications could have been utilised as an additional measure for this aspect. Such a measure could also include levels of teachers' subject expertise (for example, postgraduate qualifications) as well as teaching-specific qualifications.

There were significant concerns about scalability of the subject-level TEF exercise. If it is to be scaled up to the whole sector, the technology currently being used for the assessment process needs to be vastly improved. The QAA SharePoint system was thought to be particularly challenging, for example, making automatic password updates with lack of notification. Technological enhancements are also needed to streamline the management of conflicts of interest, ahead of scaling up.

Given the large amount of time spent on TEF in addition (for many panel members) to full-time work or study, one possibility would be the creation of sabbatical positions for academics to dedicate a sufficient amount of time to the process.

Quality of outputs

It was felt that the borderline between 'no rating' and Bronze is somewhat imprecise with respect to the number of missing data. For example, the OfS's analysis of outcomes from the first pilot initially suggested an acceptable minimum number of students as 40, but then set a more conservative threshold for the second pilot. This issue needs to be resolved if the outputs are to be of a consistently high quality.

The question of threshold achievement was discussed. If every provider entitled to access the TEF were to get a Bronze rating as a minimum, this could lead to Bronze becoming de-valued.

Reflecting on the ratings descriptors, panel members noted that that they did not always have evidence of all the qualities alluded to in these. This perhaps reflects the general feeling that the data available to assessors is very limited in relation to some of the criteria.

The aspects assessed – TQ, LE and SO – were considered at length by the BLESC panel. It was noted that the LE aspect was more likely to be awarded a Bronze rating by the panel than the other two aspects. One possible explanation for this is that LE is very varied in its make-up, including continuation metrics as well as underpinning learning infrastructures such as libraries, IT facilities and student services. By contrast, TQ, which is a more coherently defined aspect, was more likely to be awarded a Silver or Gold rating.

The suggestion was made that a new aspect could be added to recognise more explicitly 'strategic commitment to improvement'. This would have enabled panels to recognise specific steps taken, evidenced in the submission, even before their benefits had been evidenced in the outcomes data. This would have helped avert the potentially discouraging effects on providers of receiving fewer positive findings when much work for improvement was undoubtedly underway.

The question was raised of how providers would use TEF findings for marketing themselves if there were 'per aspect' ratings. If a provider's marketing focused on single aspect ratings, this could risk misleading students.

Ratings and statements of findings

Ratings

Panel members were in full agreement that ratings should be a fair reflection of the whole range of evidence before them, and that the full spectrum of qualitative and quantitative evidence provided through the submissions should be considered alongside the OfS-produced metrics. Through panel discussions, the relative merits of different kinds of evidence were discussed. There was a productive iteration between inferences that could be reasonably drawn from the initial metrics and the insights provided by the different kinds of evidence in the narrative submissions.

In many cases, consensus was reached fairly quickly by the initial trio of assessors. Where there was initial disagreement an extended discussion at stage 2, involving the sub-group of nine panel members, typically led to a resolution. A small number of submissions were less easy to resolve, and in those cases the full panel of 18 assessors plus employer and PSRB representatives was particularly important. Panel members were drawn in through the use of extra readers before an agreed resolution was reached in stage 3 discussions.

Different perspectives were shared with respect to the helpfulness or otherwise of coming to an overall Gold, Silver, or Bronze rating. As already noted, there were often differences between the different aspects of TO, LE and SO. It was felt that an overall judgement could be misleading if it masked weaknesses or supressed strengths. There was some agreement that a separate rating for each of the three aspects, without an additional overall rating, would be more accurate and therefore more helpful, not only for providers but also for students, employers and other stakeholders.

Statements of findings

The method of feedback for the second batch of submissions differed from that for the first batch, so we were able to compare the strengths and challenges of two different approaches. For the first batch, a more extended narrative statement of findings (SOF) was produced, which allowed for greater explanation of how the panel had reached its judgement. Assessors found these time-consuming to write but some preferred the approach as it was possible to convey more nuance. This was particularly helpful when there were one or two significant issues and/or areas of strength which would otherwise have become lost in a more succinct mode of feedback. Counter arguments for the narrative approach focused on the risks associated with accidental inaccuracies or unevenly expressed rationales, which could open the process up to challenge.

Students on the panel noted that the extended narrative approach to SOFs was potentially more helpful to students looking for insights into a subject's profile. However, it was also argued that students took many other factors into consideration when considering their options for study.

For the second batch, the SOFs were formulaic and short, referring briefly and explicitly to the judgements made with respect to the three aspects. These were more efficient in that they were easier and quicker to write, but it was argued by some that they might be poorly received, as no clear rationale was given for the final rating.

In their final analysis, the panel agreed a preference for a hybrid approach to SOFs, whereby a narrative approach is used but one drawing on a set range of statements to be used in varying combinations. This enhanced feedback would have the potential to encourage providers to

investigate the good practice of those doing well, as well as informing stakeholders consistently of a provider's relative strengths and weaknesses.

The panel noted that whichever approach to SOFs is adopted, the panels would need guidance materials on how to apply it consistently. Greater alignment between individual assessment templates and the SOF templates was also suggested.

Sector-wide enhancement

Panel members were keen to maximise the potential of the TEF to enhance practice across the sector. It was argued that whilst a public focus on the relative qualities of provision would in itself encourage providers to pay attention to quality review and enhancement, there were improvements that could be made to the process that would strengthen its potential for enhancement.

Some panel members argued that the profile of the metrics currently does not reward innovation but may have the effect of encouraging institutions to concentrate only on ensuring that weaker metrics improve. It was felt that the balance of the metrics is very delicate – it only takes one negative metric to rule out an initial hypothesis of a Gold rating. In some cases, what appeared in the submissions to be excellent initiatives and innovative developments did not strengthen the overall rating, as there has not been time as yet for these to make a difference to the outcomes across the three aspects.

The relationship between TEF and established routes for enhancement of practice, such as external examining, programme approval processes, quality review activities and subject interest groups, was felt to be unclear. It was broadly agreed, however, that providers receiving less than satisfactory TEF outcomes can look at the submissions of higher performing comparators and thereby identify their own areas for development.

Subject-specific observations and additional considerations

Cultural differences across disciplines were evident throughout this year's pilot in the written submissions. It was helpful to have a cross-section of disciplinary expertise on the panel, which included both wider perspectives and specialist insights.

There appeared to be some differences between the ways in which students and panel members see subjects with PSRB requirements and those without. Disciplines such as Law, Business and Education are subject to demanding scrutiny by PSRBs and it was unclear whether assessors saw approval by a PSRB as simply indicative of the TEF 'threshold' judgement of Bronze or whether it was deemed to suggest a higher standard of provision. The panel noted that there was a wide range of good practice evident across a range of providers for both Business and Management and for Law.

The complexities are perhaps greatest for Education, where a significant sub-set of provision is subject to Ofsted judgements. Ofsted applies a range of teaching-related criteria to its inspections and providers sometimes made reference to the detail of Ofsted judgements in their submissions. Panel members discussed on several occasions the relevance (if any) of Ofsted judgements to their overall understanding of the quality of provision under scrutiny. This discussion remained unresolved, and it was suggested that more discussion was needed by the OfS about the links between the TEF and PSRB accreditation.

A particular issue was raised for Law, where changes to the way the metrics measured successful further study outcomes had had a distorting effect on the SO aspect for some providers. Two of the employment metrics now only count higher level study as a successful outcome. The regulation of professional qualifications such as Legal Practice Courses means that they are often ascribed to an undergraduate level of study, and so not measured as a successful outcome by the metrics, despite the fact that such qualifications are a common and desirable graduate outcome from law courses and are required for future professional legal practice. While the panel was aware of this potential issue, it found it challenging to take into account consistently and retrospectively. Similarly, the wholesale changes about to be introduced in legal education may need to be considered in future subject-level TEF exercises.

Potential impact

The panel considered what factors were important in order for its assessments to have the most positive impact on the sector. The panel identified that the assessment process should:

- Find a balance between unstructured and over-formulaic SOFs, which should provide quality feedback
- Avoid unhelpful negative comments in the SOFs
- Be reported in such a way that it is genuinely helpful for student choice, and in ways that can be accessed by Level 3 students
- Focus providers on authentic strategic improvement, rather than 'quick fixes'
- Connect meaningfully with work being done, at the moment in parallel, with respect to access and participation plans and to the wider regulatory framework.

It was noted that TEF ratings make one contribution to a student's overall choice – there are others, such as locality, mobility, travel time and subject availability. It is important to ensure that TEF does not become a barrier to access for some, for example if a student who is only able to study in their locality were to only have access to a Bronze-rated provider.

It was broadly felt that TEF serves as a healthy balance to the Research Excellence Framework, managed by Research England.[5]

Process modifications

Members of the BLESC panel identified the following modifications to the process that would have helped their assessments:

- Through the course of the pilot, the panel recognised the challenging resource requirements of full-scale subject-level TEF for both providers and panels. The panel agreed that the value of this resource was maximised when assessment either helped maintain focus on areas that were in need of development, or where it helped showcase good practice more systematically. Developing a provider-level TEF assessment that considers subject level 'outliers', for example by focusing on the top and bottom 20 per cent

[5]See: www.ref.ac.uk/ and https://re.ukri.org/research/research-excellence-framework-ref/

of provision, would ensure that providers focus both on strategically addressing poorer provision and on promoting excellence and innovation.

- The panel were uncomfortable with the Bronze, Silver and Gold nomenclature, recognising the unintended consequences of inadvertently suggesting that, for example, Bronze equates with poor provision. These consequences could have been mitigated by using ratings of Good, Excellent and Outstanding instead. Ratings of Good, Excellent and Outstanding would also have the potential to align better with Ofsted judgements.

Medical Sciences, and Nursing and Allied Health subject panel report

Executive summary

Introduction

The Medical Sciences, and Nursing and Allied Health panel (MSNAH panel) comprised 14 academics, eight student members and three further members, one representing employers and two from professional, statutory and regulatory bodies (PSRBs). The panel met on 28 to 29 March, 1 to 3 May, and 31 May 2019 to produce subject-level ratings under the piloted model of the Teaching Excellence and Student Outcomes Framework (TEF). The panel was co-chaired by Professor Trudie Roberts, Director of the Leeds institute of Medical Education, University of Leeds and Professor Carol Hall, Director of Undergraduate Education, School of Health Science at the University of Nottingham. With Alykhan Kassam, former Pharmacy and Faculty of Life Sciences Student Representative at University of Bradford and Aaron Lowman, Former Students' Union Vice-President of Brunel University London, as deputy chairs. The panel assessed 128 submissions across eight subject areas:

- Allied health (26 submissions)
- Medical science (17 submissions)
- Medicine and dentistry (8 submissions)
- Nursing and midwifery (17 submissions)
- Pharmacology, toxicology and pharmacy (12 submissions)
- Psychology (26 submissions)
- Sport and exercise sciences (19 submissions)
- Veterinary sciences (3 submissions)

This section evaluates the subject-level pilot process and outcomes for the MSNAH subject panel.

The foregrounding of the importance of subject-level TEF was welcomed and it was acknowledged that TEF had significantly contributed to raising the status of teaching in institutions. It was recognised that sharing good practice could improve the student experience overall. However, currently there is no way for excellence within a submission to be shared across the sector.

There was concern that the subject areas Subjects Allied to Medicine and Allied Health provided a random collection of subjects. Some institutions had more than a dozen subjects in these categories, compared to other institutions with just one or two subjects, causing the panel to question the validity in providing an overall grade.

Key findings

The following key findings emerged from the process:

- Reconsider the grading system to provide a more informative system and possibly develop some form of dashboard for each subject and institution.
- For subjects with PSRBs, consider how their requirements link to the TEF criteria of excellence.
- Review the way subjects are grouped together in Allied Health and Subjects Allied to Medicine.
- Consider reviewing the weighting and usefulness of the metrics.
- More detailed feedback is preferred but should consider the impact on reviewers and the process.
- Further examine implications of small student numbers and limited data sets on the usefulness of the TEF assessment process.
- Think about how negative views on the TEF process will impact UK higher education and its relationships internationally.

Evidence, including metrics and evidence in submission

Contextual evidence and evidence in the submission

Provider-level and subject-level evidence were generally considered essential and highly valued in providing an insight into the subject being delivered. The panel recognised the critical role of subject-specific knowledge in understanding the evidence provided. Specific concern was noted with regard to the data relating to year one students who may be studying away from their place of registration. The use of the maps was varied – not all panel members used these and, when they did, they were not always found useful because providers did not refer to them in their submissions, or because the panel members did not feel confident to interpret this evidence.

Key points

There was value in provider-level and subject-level contextual data, particularly in some subjects e.g. Pharmacology, Toxicology and Pharmacy, which contributed to in-depth discussions and changes in ratings as panel members moved through each step of the assessment process.

Panel members would have liked to have direct access to student declarations, which they felt were an important component.

It can be difficult to make decisions because the metrics are so high amongst some subjects. Therefore, the narrative plays a big part in decision making.

Subject knowledge of the panel is important when comparing the individual provider subject metrics with the data for the subject across all providers (including data outside the TEF subject-level pilot subjects).

If maps are to be used, panel members and providers need further guidance in interpreting this information.

Criteria

The change to the criterion to divide student engagement into learning (TQ1) and student partnership (TQ5) was positively perceived as it offered a means for distinguishing excellent practice. It was noted that a provider may have a didactic approach to delivery, such that students may be highly engaged and achieve outstanding outcomes but with little influence on programme infrastructure. TQ5 distinguishes those providers that take a more collegiate and responsive approach to learning and teaching.

Views on the metrics

The panel found the metrics adequate to interpret, and the systematic nature of the process satisfactory, but noted a number of concerns. These concerns increased the burden on panel members compared with previous years (which will be addressed in the 'Assessment process' section below). This section focuses on evaluation of the metrics weightings and the structure and content of the workbook. There was acknowledgment that the TEF exercise remained weighted towards the metric outcomes and that greater emphasis needed to be placed upon the holistic submission.

Weighting within the metrics

The panel questioned whether the current system of weighting of the metrics was effective in assessing medicine and health sciences subjects. Providers could be identified as Gold initially even with poor NSS metrics, thus meaning that the initial hypothesis did not reflect the wider student experience. Greater weighting could be placed upon NSS metrics. Subject specialists observed that this could be problematic, however, with regulated courses and those including a high inclusion of practice placement time. Nonetheless, the provision of such courses include accreditation where the partnership between placement provider and higher education provider is assessed, so potential imbalance might be considered a threshold issue.

The panel questioned whether the continuation metric was weighted too heavily, recognising that values-based recruitment can influence continuation. Professional courses such as nursing or medicine may have higher voluntary continuation, but progression is restricted by professional fitness and competence in order to progress into future practice, which may force attrition. The heavy weighting was also recognised as offering potential for providers to gain advantage through 'gaming' the system. The panel noted that, in some short programmes, the continuation metric was actually showing completion rather than continuation.

As many professional courses were included within the CAH2 subjects being assessed, there was concern that the employability metric was weighted too heavily. In courses not directly leading to employment, it was noted that this may be disadvantageous. For example, there was little shift from the initial hypothesis to final holistic rating for Psychology subjects and the subject specialist

noted the possibility of issues around employability. Employment metrics could be influenced by students using their course as a stepping stone to another degree (mostly for Medicine). This could negatively impact highly skilled employment or further study metrics because Medicine is not counted as further study, yet students are able to progress to a Masters equivalent course through this step.

Robustness of Longitudinal Educational Outcomes (LEO) data as core metrics

There was concern regarding the LEO data and its applicability in respect of new programmes. In some cases, courses had only run for three years and had insufficient data to be assessed. In other cases, courses within the LEO data had been superseded by others within the subject and were no longer running. The use of LEO data does not fully support TEF to enable an agile curriculum and new programmes. Additionally, the highly skilled employment data for sports programmes can be misleading as the definition of highly skilled employment poses significant questions in this occupation.

Key points

- Metrics are too heavily weighted to outcomes – a greater focus on teaching excellence and student experience is needed.
- Initial hypothesis in TEF is data driven, but regulation is more subjective, although viewed by many as being a higher standard.
- Some new programmes had insufficient data to be assessed. Where LEO data was provided for other courses, the data was often out of date and did not relate to the course currently running.

Workbook structure and content

The colour coding of metrics within the workbooks was identified as offering a useful indicator along with the inclusion of flags, absolutes, material difference and Z-scores – all of which were identified as helpful in offering initial insights or raising questions. However, there were a number of reservations, mostly relating to the extent to which the indicators could influence panel members in their decisions. It is important to note that panel members questioned whether the flagging colours in the metric workbooks could limit decisions away from the initial hypothesis. There were mixed opinions regarding the use of two confidence levels for statistical significance (Z-scores of +/-1.65 and 1.9). Some panel members found these useful to observe trends, whereas others identified them as invalid. A further question was raised over the statistical significance of metric flags, which highlighted panel members' lack of clarity when completing assessments. The metric workbooks contained a large number of multiple comparisons and interpreting these was thought to not have been fully addressed.

Key points

- Visual marking of metrics was good – e.g. an array of green gave a clear snapshot.
- Colour coding of statistically significant metrics was also useful, e.g. flags, absolutes, Z-scores.
- Colours could provide the wrong initial impression and may limit the ability to form a final hypothesis. However, colours could also highlight areas for further investigation.
- The +/-1.65 and 1.9 Z-scores should be used with caution or removed, given their limited statistical validity.

Small data and non-reportable or missing metrics

The validity of small reported numbers was debated, particularly where this led to unreportable data in the split metrics. The issue was intensified when unrelated small programmes were combined within a subject submission, and where unreportable metrics were produced, resulting in great difficulty in reaching robust decisions. In a number of cases where there was insufficient data within the metrics, narrative submissions were found to be insufficient to make sound judgements. 'No rating' decisions were made where a default to Silver or Bronze may have been given last year. It was felt that some subjects graded 'no rating' were probably doing an outstanding job for their constituencies but reaching a rating based on the narrative alone due to the absence of supportive metric evidence raised caution across the panel.

The panel considered a minimum threshold number for inclusion in review but this was difficult. The view was that consideration should be given as to whether a threshold number could be decided on before the submission. Analysis of the panel's use of 'no rating' suggests that 40 students might be an appropriate figure, but only if the subject has reportable metrics covering all three aspects. If core metrics show significance can be achieved, it should be possible to have confidence in an assessment even if they are non-reportable in splits, but this was difficult for panel members. The panel noted that some aspects could technically be rated (even if not all could be and even if an overall rating was not possible), and this could offer information to providers. Finally, the panel expressed concern that 'no rating' could be viewed negatively, whereas in fact it meant there was not enough data to enable a rating to be made.

Key points

- Judgements of ‘no rating’ are problematic and could give a negative impression.
- Given the importance of the metrics in providing the initial hypothesis, missing or unreportable data are problematic.
- Low numbers in the widening participation metrics at subject level made it difficult to draw meaningful conclusions. This analysis may be more robust at provider level.
- It would be simpler to have a threshold ruling on small data – e.g. below 40 students.

- Small and new programmes had limited data so scored no flags, giving them a default position of Silver, but this was essentially formed on the basis of no metrics.
- Providers referred to their own metrics, but within the context of the submission it was difficult to assess the validity of claims – arguably different criteria are being applied.

The assessment process

Robustness of the process

The whole panel was in consensus that robust judgements were made. Each case was well considered and outcomes embraced a range of thoughts and views. Panel members valued each other's diverse perspectives, and especially student members, who the panel identified as 'excellent', offering critical consideration and a different viewpoint. Transparency was valued and seen to support due diligence, although some panel members felt that this could have been achieved through a more thorough calibration and use of the trios. The potential for inequality of opinions by different groups (e.g. students versus academics) was raised and while the students themselves did not identify feeling undervalued, they did find it more difficult to comment and assess courses where they did not have specialist knowledge.

Representativeness on the panel was identified as influencing robustness where a large subject had no representation (e.g. Dentistry) or participation by a sector was limited (e.g. further education colleges (FECs)), or where a subject was divided between CAH classifications. New panel members (including students) identified a steep learning curve, with experience being developed 'on the job' which could influence the robustness of individual contributions. Training videos were praised and considered a great improvement but thought should be given to buddying new panel members with experienced ones.

Key points

- Academics with FEC experience and FEC student panel members should be identified.
- New panel members would benefit from buddying up with more experienced panel members.
- Representativeness of the panel is important.
- Overall, panel outcome judgements were robust.

Panel consistency

Review of the cross-match cases and the panel outcomes data revealed similarity between the two MSNAH panels. However, it was perceived that some panel members gave more weight to some metrics than others during discussions, thus affecting consistency. The panel was concerned about whether the NSS metrics were correctly weighted. Further, the panel identified that the assumption that every student wants the same outcome from their degree should be questioned. Nonetheless, it is pertinent to question whether these individual perceptions influenced consistency between panel members even if the whole panel outcome was finally consistent.

Key points

- Cross-check cases showed similar judgements across groups, thereby implying consistency.
- Differences in subject metrics (e.g. Medicine and Dentistry have high continuation and employment metrics across the sector) mean it can be difficult to differentiate them, particularly when thinking about the weighting of NSS metrics (the main differentiator in these cases) in comparison to the weighting of outcomes.

The revised CAH classification

The Medical Sciences classification enabled a more consistent range of subjects to be assessed this year, although panel members still questioned whether an overall rating for an area that included up to 11 subjects would help students. Medicine and Dentistry are usually large courses and some panel members questioned the rationale for combining these subjects.

Allied Health continues to encompass a disparate range of subjects with differing expectations and employment and career trajectories (e.g. paramedic science and counselling). Furthermore, some professional courses became 'invisible' (e.g. operating department practitioner). Overall holistic judgements could not always reflect all courses included and thus may be a poor indicator of TQ.

Key points

- Consider separating Medicine and Dentistry into separate subjects.
- Concern regarding 'invisible' professional courses.
- Reflect on other ways to provide a sub-judgement for Allied Health subjects when course outcomes can be clearly assessed.

Effectiveness of widening participation and interdisciplinary liaisons, employer and PSRB roles

Panel members had differing views on the effectiveness of the specialist panel roles, but all agreed that individuals undertaking the roles was valuable. Some pragmatic suggestions were made.

Widening participation (WP)

The WP role was identified as embedded within the process – the panel was aware of this agenda and minority groups were considered as part of the holistic judgements made during evaluations. The spreadsheet compiled for this purpose was challenging as it was too big to scroll through and across and keep track of the column headings. Drawing attention to vertical splits for particular student characteristics as each assessment was discussed and referring to the provider-level geographic context and student demographics compared to the subject-level student demographics, etc. worked better. It would be helpful to have this information summarised.

Interdisciplinary liaison

Interdisciplinary liaison roles proved difficult – some panel members felt that this role was not effective in achieving its aims, while others felt more could be made of them. Submissions could add comments on other courses that are split with the subject being reviewed; the percentage split; and whether this is normal across the sector or unusual to that particular provider.

PSRB and employers

PSRB representatives were highly valued by panel members for their contribution. However, PSRB representatives did not always identify this value, and the clarity of the role and sense of purpose was recognised as being critically important. The panel felt that the influence of PSRB expectations on the evaluative process needed clarification. In a small number of cases, student panel members expressed that they thought a subject looked outstanding while academics highlighted it as common practice or a PSRB requirement. While the TEF process must be credible across professional communities served by the delivery of graduates, there was concern that PSRBs and TEF were measuring and weighting different elements of professional courses. There were questions raised as to how excellent practice was to be taken into consideration when making a rating for something which was considered a standard expectation by the PSRB. Greater clarity to support consistency would be valuable in ensuring a level playing field for PRSB and non-PSRB monitored courses. Concern was noted regarding a lack of employer representation and limitations in panel diversity.

Key points

- Panel members are aware of WP needs and addressed these in their assessments.
- Subject specialists and employers bring their expectations to the discussion but non-specialists can make a valuable ‘outside’ viewpoint.
- The PSRB role requires clarification in terms of expectations and influence of standards.

Workload and burden

The workload was considerable and this was particularly identified by student panel members. More timely availability of the paperwork for meetings was requested, as well as better and more consistent use of SharePoint to minimise time spent looking for information. Viewing metrics on some computers was challenging but easier with access to two screens. Not all panel members had access to this technology, students in particular. Changes made throughout the process were identified as additional burdens leading to lower confidence in the evaluations regarding consistency between the batches.

Key points

- The workload was burdensome, especially for students. This includes in-process TEF changes.

- Appropriate reading time should be allocated to take account of differing reading times required by panel members.
- Panel members have built up expertise which should not be lost for future TEF activities.
- Getting to grips with the metrics was challenging. Panel members did not always know how to manage ‘greyed out’ statistics.

Reflections on the robustness of the trios and nines

Training was identified positively and was supportive to the robustness of the process although the 'Kick Off' event in London was not considered helpful for new participants due to being large and overwhelming. The trios worked well and provided the breadth and depth of assessment required.

Reporting on outcomes and coming to a deliberative final outcome in nines on day one worked well and ensured consistency in deliberation and outcome. Some felt the day was too long and having the more complex cases at the end did not ensure efficient deliberation. The second day was felt to repeat the first day’s deliberation. However, there was value in discussing more complex cases with the wider group and agreeing a consensus approach.

Key points

- Student panel members were felt to be equal members of the panel and made a positive contribution.
- When additional readers are requested, information should be provided as to why the request has been made, with specific questions given to additional readers to address. This would make the process more efficient and avoid repetition of discussions at stage 2 and stage 3.
- It is difficult for panel members who have not read the submission to make a valuable contribution during whole-panel discussions.
- The lead reviewer role was helpful in speeding up the process.
- It was difficult for panel members to remain focused when their allocated submissions were at the end of the day.
- The process could be made more efficient by reviewing some submissions once instead of twice across the stages e.g. clear-cut cases from the initial assessment stage, which resulted in an agreed rating. The remaining time in the final stage of assessment could focus on whole-panel discussion for difficult/borderline cases.

Ratings and statements of findings

Fitness for purpose of rating scheme of Gold, Silver, Bronze

The current ratings were not deemed fit for purpose by the panel, as they did not fully communicate the general excellence in the sector. The panel felt that this shortcoming had potential detriment to the sector's international standing and students' understanding of providers' quality. Discussion of borderline cases is an important part of the process. In relation to the panel outcomes, stage 1 has a five-point scale including borderlines, which has to be reduced to a three-point scale, so the borderlines have to be moved. In the MSNAH panel, written submissions helped borderline Gold/Silver initial hypothesis cases move to Gold at the final rating. However, more borderline Silver/Bronze moved to Bronze, indicating a weaker written submission. The second batch of assessment had more Gold and less Bronze than the first batch of assessment. This was due in part to different starting points at the initial hypothesis and a greater number of more strongly written submissions. Nonetheless this also may relate to panel differences as they became more familiar with the process. Going forward, consideration needs to be given to the composition of panels and the 'churn' with respect to the numbers of new and experienced assessors.

System for awarding for each of the three aspects of the case

Panel members liked providing responses to the three aspects of quality as they could more easily bring out some areas of provision which were to be commended. They did identify though that per aspect ratings can differ in different applications but lead to the same overall rating for subjects. It was suggested that this may confuse providers, especially if SOFs do not provide enough information around panel decisions. This was especially identified where submission and metrics differed significantly. The process for the second batch was more logical and rating each aspect helped with breaking things down prompting a more thorough analysis.

Key points

- More cross-checking is required. In particular, moderating and cross-checking clear Gold, Silver and Bronze cases could be useful in determining examples of ratings at a mid-point of the rating criteria.
- A finer-grained numerical rating may work better to indicate a range of excellence.
- Per aspect ratings should be retained as these are useful for giving information to providers and students, as well as for making a more detailed analysis.

Statement of findings

The principle of the statements of findings (SOFs) in adding an opportunity to share more useful information to students and providers was welcomed. However, the SOF needs to be timely and done at the assessment stage – it was found challenging for lead reviewers to contribute meaningfully and lead the discussion as well as take notes in the meeting for the SOF. The more descriptive (first batch) SOFs provided more support for providers who wish to use the feedback to enhance their educational provision but they do take considerably more time to produce, especially

if they are to be of high quality. Additionally, they may be more open to challenge. A future SOF should define who it is for and it needs to be equitable – the scalability of producing a detailed SOF should also be considered.

First and second batches

The template for the first batch and the suggested wording were useful, although panel members found it hard and time consuming to create unique and insightful responses of benefit to providers and to students. Chairs and TEF officers reviewing the first batch of SOFs following submission by panel members found variability. The editing process was lengthy before they could be signed off. The second batch of SOFs was deemed easier and less burdensome on an assessor in terms of time and effort. However, panel members commented that it would be difficult to know how valuable they were to the providers. Nonetheless, these did seem useful when an assessment was borderline.

Key points

- SOFs in the second batch were less burdensome, and may be potentially less useful to providers. They are more consistent than the first batch.
- Use of student declarations would be helpful for the panel.
- Issues raised in SOFs may suggest to the provider that these are the only areas that they need to address – it can be difficult to write SOFs both for the institution and to inform student choice. However, the SOF text does need to be directly related to the criteria.
- A list of useable and standard template statements in an electronic form provided by the OfS would be useful – perhaps SOFs could be completed using drop-down boxes during assessment meetings to avoid errors, increase accuracy and reduce burden on assessors.

Sector-wide enhancement

During the assessment process, panel members identified many examples of excellent practice across the sector and much evidence that could help situate claims by providers in the context of the general subject requirements. Subject-level assessment offered considerable information, including excellent production of narrative submission, excellent resources and excellent practices.

Subject-specific observations and additional considerations

From a subject-specific perspective, there were concerns over the weighting of NSS metrics against the employment metrics for professional programmes; issues with small courses; and ambiguity in the role and participation of employers, subject specialists and PSRBs. Some panel members felt that assessments should be completed anonymously to avoid academics and students bringing their prejudices, even given the acceptance of panel member integrity. Blind data would make a much more defendable process even though the process of blinding applications would be problematic.

Additional observations

- The second batch of assessments included larger providers, and submissions appeared more developed in some instances, leading the panel to comment that batches could be more mixed.
- There was some concern that the OfS could potentially be subjected to freedom of information requests for SOFs and the implications of this for panel members needs to be considered.
- The majority of panel members felt that chairing and deputy-chairing arrangements worked well. Nevertheless, some suggested that more externality was required.

Potential impact

The panel felt that TEF at provider level and subject level has done an excellent job of foregrounding the importance of teaching within the higher education sector, offering evidence and a need for subject-level providers to reflect over their metrics, as well as consider their own strengths and areas for development. The assessment process, while still in refinement for subject level, shows potential for subjects in understanding their sector more fully; driving enhancement and improving key areas of concern including WP and black, Asian and minority ethnic attainment.

Information for students

TEF needs to do more in its role as a student information tool, to clearly identify areas of good practice and highlight them into a digestible form for potential applicants. Panel members noted that it is difficult to talk about students as one large group, when they are not a homogenous set. More student input was recommended as a limited involvement in submissions and half-weighted NSS data was not considered sufficient in highlighting the student experience. Also, perhaps students should be able to view the ‘per aspect’ ratings with relevant supporting information.

Unintended consequences

Ratings may affect the desirability for international students to study in UK higher education; particularly to study a subject which is not rated Gold, thus causing potential harm to the reputation and income of UK higher education. Subject-level TEF could also stimulate institutions to examine the range of their provision and target this to CAH2 areas. This could lead to a reduction in interdisciplinary and minority subject provision at a time when the employment context could well favour more interdisciplinary provision. TEF data could be used inappropriately to draw up league tables that would not reflect some institutions’ high quality educational provision that performs an important local educational function but does not produce graduates who are highly remunerated. Finally, comparing further education colleges with internationally renowned Russell Group universities seems unreasonable due to the disparity of funding between the two types of institutions.

Key points

- All institutions will be concerned with their market position – their subject-level TEF rating being another driver.
- All institutions will aspire to Gold, although many will not have the investment potential in the current climate to make a full response. Therefore, institutions are likely to refine their offer to achieve best outcomes in TEF – this could include the deletion of poorly rated programmes.
- Gold-rated subjects will wish to maintain their competitive advantage.

Natural Sciences, and Engineering and Technology subject panel report

Executive summary

Introduction

The Natural Sciences, Engineering and Technology panel (NSET panel) comprised 14 academics, seven student members and four further members, two representing employers and two professional, statutory and regulatory bodies (PSRBs). The panel met on 28 to 29 March, 1 to 3 May and 31 May 2019 to produce subject-level ratings under the piloted model of the Teaching Excellence and Student Outcomes Framework (TEF). The panel was co-chaired by Professor Sue Rigby, Vice-Chancellor of Bath Spa University and Professor Nick Lievan – Professor of Aircraft Dynamics at the University of Bristol. With Lauren Marks, former Students' Union President, Education Officer and course representative, Aberystwyth University, and Lewis Cleminson, Students' Union Education Officer of Southampton Solent University, as deputy chairs. The panel assessed 139 submissions across eight subject areas:

- Bioscience (21 submissions)
- Chemistry (13 submissions)
- Computing (34 submissions)
- Engineering (29 submissions)
- General, applied and forensic science (12 submissions)
- Materials and technology (9 submissions)
- Mathematical sciences (14 submissions)
- Physical and astronomy (7 submissions)

This section provides the key observations from the NSET panel for the second phase of the TEF subject pilot in 2018-19. The commentary has been derived from panel discussions and from written feedback from panel members, having reflected on this year's cycle.

Key findings

- 'Rigour and stretch' was a key criterion for the TEF assessment and currently there is no metric which helps inform the assessment. The inclusion of the National Student Survey metric Q04 'course is intellectually stimulating' could be considered.
- Negative performance in one metric at step 1b should not preclude a subject being awarded a Gold TEF rating. The panel felt the formulaic method for arriving at a starting point rating was too limiting in terms of the number of subjects which started at Gold.
- The classification of subjects for assessment by panels using the Higher Education Classification of Subjects (HECoS) Common Aggregation Hierarchy (CAH2) requires a sense-check from higher education providers. The NSET panel felt that a number of its cases would have been more appropriately assessed by other subject panels. It also felt that

providers should be allowed to make the case for individual subjects being assessed by alternate panels which are a better fit to the curriculum.

- Although continuation (the proportion of entrants in a given academic year who continue studying in UK higher education) is a useful indicator of the quality of learning environment, completion (i.e. graduation) would be a fairer reflection of the whole student experience.
- Longitudinal Educational Outcomes (LEO) data did not command the confidence of the panel, not least because of their historic nature and uncertain relevance to current students. However, the above average median earnings metric was used as an effective proxy (and should be retained in the core metrics).
- The feedback on assessment provided through the statements of findings (SOFs) should be sufficiently descriptive to encourage enhancement in the sector. The panel recognised there would be significant risk of challenge with more detailed commentary and so suggested that this feedback should be 'in confidence'.
- Confidence in the assessment process grew with experience. In future TEF assessment models, consideration needs to be given to both increased early stage training and the recruitment of panel members with appropriate experience and expertise to come to sound judgements.
- For assessment at subject level, it was important to have assessors with relevant expertise in the subject.
- The group of three initial assessors for each case should have subject-level expertise within the discipline.
- Benchmarking of metrics data ensures higher education providers are treated fairly in the assessment; however, it may be confusing to applicants when deciding between providers.

Evidence, including metrics and evidence in submission

Metrics

Metrics weighting: The panel felt that different weighting of metrics was unnecessary. In particular, it was felt that the double weighting of the continuation metric in formulaically determining the step 1a initial hypothesis was overstating its importance. Furthermore, completion would be a fairer measure of the student journey and would provide a more holistic perspective of progression through the academic programme.

Having five metrics derived from the NSS placed too much importance on this survey, especially since these indicators are closely correlated. Overall, weighting of metrics for the step 1a initial hypothesis should not be taken as an indicator of the weight given to each individual aspect of quality. For example, TQ metrics have a weight of 1.5, whereas LE and SO metrics each have a weight of 3.0.

Criteria and metrics: The metrics do not directly map onto the TEF criteria, for example rigour and stretch (TQ3), and scholarship, research and professional practice (LE2). Although rigour and

stretch is a key measure, the panel had to try repeatedly to interpret the narrative in the absence of a core metric. The inclusion of NSS Q04 ‘do I find my course intellectually stimulating’ and, similarly, inclusion of measures relating to valuing teaching (TQ2) would indicate provider-level commitment. Alternative metrics for some of these criteria could be considered, e.g. indicators of teaching qualifications, and a wider interpretation of student outcomes beyond employment and graduate salaries (although above median earnings was used consistently).

Metrics splits: Breaking metrics data down by different characteristics (metrics splits), e.g. ethnicity, caused issues when there were small numbers of students in the group in question as this led to data being suppressed. In particular, the panel felt that providing metric indicators for different years of study was unnecessary as the process did not allow for consideration of trends and trajectory, though it was felt that being able to consider such trends in their assessment would be useful. The panel also acknowledged that the metrics splits were important for consideration of widening participation (WP).

Comments on specific metrics

Longitudinal Educational Outcomes (LEO) data is historic and it is problematic that it is not benchmarked regionally, as employment outcomes are known to vary significantly across the UK. In particular, the above median earnings metric often looked out-of-step with the other employment metrics, although these were considered to be more robust. These arguments were made in a number of submissions.

Submissions

The panel noted that the provider-level summary statements were very helpful, although they were not always well aligned with the subject submissions. This may have been a result of the timings of the submissions during the pilot exercise.

The panel felt that student submissions at subject level might be a useful addition to the assessment process but were cautious about how much value such submissions might add to the assessment, and about the burden of producing such documents. For example, careful consideration would need to be given to timing of submissions in order to avoid exams and major coursework submissions.

The panel had concerns about data protection issues with individuals being named in some submissions (e.g. external examiners, or students providing quotes).

The panel felt that additional pages at subject level to cover cases where there is a significant proportion of part-time provision would be beneficial.

Similarly, where a submission needs to cover a wide range of courses, extra pages might be appropriate. There were a few subjects within NSET where this was a particular issue (Biosciences; Materials and Technology; and General, Applied and Forensic Sciences).

Submissions were variable in quality and usefulness. Some providers were able to improve their rating by having written a strong narrative. Others, perhaps no less strong in terms of quality of education provided, did not improve their rating. The panel suggested that 'good practice guidance' is given to providers about how to write a strong submission.

The assessment process

The formulaic approach used to derive the initial hypothesis seldom resulted in a Gold starting point, largely because the existence of a single negative flag immediately ruled out this possibility.

The panel felt that some form of initial discussion between the three individual panel members prior to the group discussions would be a more efficient use of time at stage 2 of the assessment. Ideally this would be face-to-face. Panel members could highlight if there were need for additional reviewers at an earlier stage which could lead to a more efficient process.

The classification of subjects for assessment by panels using the Higher Education Classification of Subjects (HECoS) Common Aggregation Hierarchy (CAH2) requires a sense-check from higher education providers. The NSET panel felt a number of its cases would have been more appropriately assessed by other subject panels. Consideration should be given to allowing higher education providers to make the case for allocating a subject level submission to the appropriate panel.

While subject specialism is important, reviewers were able to read outside of their direct area of expertise. In some cases, there was an imperfect match of reviewer expertise and in a small number of cases (e.g. in Materials and Technology, and in General, Applied and Forensic Sciences), the panel would have liked to consult another panel with more appropriate expertise. Perhaps using a cross-referral mechanism to consult a specialist advisor akin to Research Excellence Framework (REF) processes[6]. Ideally, each group of individual reviewers considering a subject submission would include discipline specific expertise for the system to be credible and robust, for example Mathematics or Chemistry expertise and not just Natural Sciences. However, the panel recognised this would involve considerable constraints in distributing the assessments to reviewers.

During the second batch of assessments, panels were asked to give ratings to the different aspects of quality for a subject (TQ, LE and SO) in addition to an overall holistic judgement for that subject. Reaching these per aspect ratings was not problematic for the panel and, in many cases, aided the panel in reaching an overall holistic judgement. Specifically, the panel felt more comfortable settling on an overall rating of Gold, Silver or Bronze if they were able to reflect a more nuanced picture with the per aspect ratings; it was especially useful to be able to make borderline judgements on the aspects of quality.

The panel considered submissions sequentially by provider which helped identify subject performance relative to the institution's subjects that fall within the subject group assessed by the panel. This contextual approach worked well, with time allocated at the end to consider the distribution within individual disciplines looking across all providers.

The panel saw the value in 'following the process in reverse'. In a handful of cases, where an additional reader was assigned, this panel member began by looking at the submission, rather than with the metrics. This gave a fresh perspective, and could help to recognise the importance of the submissions and avoid a judgement driven too strongly by metrics data alone. In addition,

[6] "In cases where, in the sub-panel's opinion, the sub-panel and its appointed assessors do not have the required expertise to assess specific parts of submissions, those parts of submissions may be cross-referred to other sub-panels for advice." REF 2019 panel criteria and working methods, paragraph 399. See: www.ref.ac.uk/ and www.ref.ac.uk/submission-system/.

training should not just focus on the metrics, but should cover evidence in submissions in much greater detail. The threat of 'metrics capture' could be avoided by training sessions which have more emphasis on reading and analysing narratives, and on generating a shared understanding of how to handle 'difficult' cases, and less on using and interpreting metrics.

As far as possible, the process should try to balance the cases examined in different batches, putting together similar profiles of cases with a similar mix of subjects and provider types starting at Bronze, Silver and Gold, and borderline ratings at step 1a. This would allow for simpler calibration between assessment batches and would be especially important if there were more than two batches.

For approximately 10 per cent of cases, the panel felt there was insufficient evidence to award a rating and therefore the minimum number threshold for a subject to be eligible for assessment was too low. The panel spent a disproportionate amount of time discussing these cases.

More training and larger-scale, more detailed, calibration exercises would have been necessary to ensure a robust process was workable over multiple breakout groups working across multiple batches of assessment.

Employer and PSRB representatives were key members of the panel. They felt their role could be much closer to the other panel members. For example, they felt that concentrating on the submission, and not engaging as fully as other panel members with metrics, limited their ability to contribute to group discussions.

A number of process efficiencies arose through the assessment process and these should be considered in future TEF exercises:

- During full panel discussions at stage 3, cases with unanimous agreement from the previous group discussions at stage 2 were quickly signed off by the panel, allowing more time for detailed discussions where necessary. The panel questioned whether this sign-off was necessary if it had overall confidence in the judgements of the sub-groups.

- Allocation of additional reviewers was used consistently on all cases which remained at borderline TEF ratings after group discussions during stage 2 of assessment. In particular, it was helpful to have more than one additional reader.

- A number of cases were reviewed by both sub-groups of the panel. These cross-check cases were helpful to calibrate between the groups and the panel felt that including more such cases would derive even more benefits.

Some aspects of the assessment which were applied at provider level might be worth considering at subject level:

- **Subject-level student declarations**: The panel felt that some submissions may have assumed that the student voice was captured in the student declaration and therefore did not need to be covered in the subject submissions.

- **Additional pages**: Additional pages at subject level to cover cases where there is a significant proportion of part-time provision.

There was some confusion in the student declarations between showing engagement in the TEF submission and engagement in their learning. The latter is far more important and should be the emphasis of the student submission.

During both assessment meetings, the student deputy chairs moved between the assessment groups, attending sessions with each. This was a benefit to the process as the deputy chairs were able to provide observations across both assessment groups.

Ratings and statements of findings

Ratings

The panel was conscious of how the TEF ratings are perceived externally especially from an international standpoint. In particular, it should be clear that Bronze is a positive award and that where the panel felt there was insufficient evidence to reach a robust judgement and therefore returned a result of 'no rating', it should also not be interpreted as a negative result. However, the panel felt that star ratings, akin to the REF, or descriptive ratings such as those used by Ofsted (Outstanding, Good, Requires Improvement and Inadequate) might be more indicative of the panel's judgements and suppress the negative connotations associated with a Bronze outcome.

The panel observed that overall ratings profiles were reasonably balanced between subjects. Variation between subjects could be rationalised by sample effects, for example where the panel had only assessed a relatively small number of cases in a particular subject area, or where the sample of cases for a subject which were examined by the panel were not representative of the subject within the higher education sector as a whole. The panel did highlight a number of observations of the data:

- Compared to other subjects, there did not appear to be much movement upwards from the formulaic step 1a initial hypothesis in Mathematical Sciences in contrast with Chemistry. Both subjects started with no clear Gold at step 1a, but Chemistry ended with significantly more Gold ratings when the panel reached its final holistic judgements.

- There were relatively few engineering cases which had a formulaic starting point of Gold at step 1a, with the panel concluding this could not be entirely down to sampling effects due to the statistically significant number of cases (29) assessed by the panel. The panel observed that generally high performance for Engineering in the sector as a whole may have led to high benchmarks, in particular for SO, which made it difficult for subjects to perform significantly above the benchmark and receive a positive flag. Examination of sector data of metrics lent some weight to this argument; however the same argument could be made for other NSET subjects, in particular Physics and Astronomy, where a reasonable amount of Gold ratings were seen at step 1a.

The panel felt that if cases were clear Gold, clear Bronze, or definite Silver, based on the metrics, then the submissions had little chance to influence the assessment and questioned the value in considering these cases. Examination of the data showed that:

- of 19 cases with a Bronze rating at step 1a, three were awarded a final rating of Silver

- of 68 cases starting with a Silver rating, five were awarded Bronze and 10 were awarded Gold

- of 14 cases which started with a Gold rating, one was awarded Silver.

These outcomes confirmed that movement from the initial hypothesis not only occurred in cases with an initial borderline rating at step 1a, but also in cases that were more centrally located within a band.

The Silver category is very broad and because the majority of cases are awarded Silver, this is seen as the norm and anything below that rating is seen as below average. The issue therefore might be in the formula for arriving at the step 1a initial hypothesis which influences the overall profile of ratings.

Ratings given to the different aspects of quality indicated that SO was more often given a different rating from the overall holistic judgement, for both higher and lower ratings. The panel felt that this was not surprising as employment outcomes played a major part in their discussions, even with the misgivings about LEO data referred to earlier.

The data on interdisciplinary provision largely reinforced the panel's knowledge of expected combinations of subjects. The panel was interested in where NSET subjects combined with those from other subject panels, but again these were largely as expected e.g. Biosciences combining with subjects from Medical Sciences, Nursing and Allied Health, and Computing, Engineering and Mathematical Sciences all combining with subjects from Business and Management, and Education and Social Care. Accordingly, consideration should be given to a generalist panel to examine provision where there is no natural fit for a discipline.

The observation was made that the panel spent the majority of its time resolving borderline cases such as those with an initial hypothesis of Bronze/Silver or Silver/Gold. If the process allowed borderline ratings (on a five-point scale), then potentially the process could be streamlined. There was a concern that more ratings would just lead to more borderlines, but this might be somewhat mitigated by retaining rating descriptors for Bronze, Silver and Gold, and using Bronze/Silver and Silver/Gold for genuinely borderline judgements.

How participant providers are informed of their ratings for the different aspects of quality needs to be carefully messaged, due to the perceived inconsistencies in the results and because the overall rating is a holistic judgement, not based on a formula combining the per aspect ratings. Some examples of this could be:

- Ratings for individual aspects of quality, but no rating overall
- Ratings provided for individual aspects of quality where there are missing metrics for that aspect, but mitigations are made in the submission
- Explaining the relationship between the overall rating and for individual aspects of quality to mitigate for appearing to be inconsistent, for example when a subject has two aspects rated as Silver, and one rated as Bronze, but with an overall Bronze judgement.

Statements of findings

Narrative statements of findings (SOFs), produced during the first batch of assessments, are probably of more use to providers, but the panel felt there was a significant burden in production of these documents, and acknowledged that significant risk of challenge to the process could arise

from this option. Capturing the panel discussions in a way that is suitable for publication was challenging, as individual panel members write differently and the panel was concerned that an individual assessor could struggle to report the collective view of the panel. The process could be adapted to improve these issues by allowing a mechanism for other reviewers to check that content of SOFs was a fair reflection of the collective view. However, better training on writing SOFs should be provided to panel members if this approach is adopted for a full-scale exercise.

In addition, if these are to be public-facing documents, then use of language needs to be reconsidered. Discussing an 'initial hypothesis' is not really accessible to potential applicants.

The formulaic SOFs which were produced from the assessment of material from the second batch of assessments did not provide additional information other than the ratings for the three aspects of quality. The production of this format of SOF could be automated which would be minimal burden to panel members.

The panel felt it would appreciate an opportunity to give feedback to providers privately in a form that would not to be published. Such feedback could be given as highlighting good practice and areas for development and not give a detailed account of how the panel reached any particular judgements. The panel felt this would be a good opportunity to drive enhancement in the sector. However, the panel acknowledged that there could still be concerns about such feedback giving grounds for appeal.

In addition, feedback provided on good practice for writing submissions would be useful to the sector as a whole.

Subject-specific observations and additional considerations

Some subjects proved more challenging than others to review due to their cross-disciplinary nature. Examples were Materials and Technology, Biosciences and General, Applied and Forensic Sciences.

The panel found a number of courses difficult to review, such as HNCs and professional courses, and questioned whether they should be in scope for subject-level TEF.

While the panel felt that subject expertise was necessary for the robustness of the assessment process, further clarity on the role of the subject specialist reviewer would be helpful. The panel was concerned that, in some cases, the match of subject expertise was not always close.

Potential impact

Usefulness to students and applicants

Overall, the panel felt that the outcomes of subject-level assessment would be useful to students. Student panel members particularly felt that more information to inform student choice would always be 'a good thing', and of particular interest to prospective international students. The ratings for the different aspects of quality might be especially useful in providing a helpful granularity to the ratings. However, certain aspects of the process, such as benchmarking, might be difficult to interpret for prospective students.

Scalability

The panel felt that more training and larger-scale calibration exercises would be necessary to bring a larger panel (necessary for a full-scale subject-level TEF exercise) up to speed with new processes required for a revised assessment.

'Cross-check' cases were extremely useful to calibrate between the two groups of panel members. The panel felt that a scaled-up model would benefit from an increased number of such cross-check cases.

Recruitment of panel members with suitable experience and expertise is expected to be a challenge and has to be combined with significant training. There was no doubt, even with a relatively small and well-matched panel, that decisions became more confident and robust over the duration of the two-year pilot. The NSET panels would also have benefited from being more diverse and this should be an area of focus when recruiting panels in the future.

Conclusion

The key findings for the NSET panel are summarised in the Executive summary and will not be repeated here. However, in writing this report, the panel felt that it was trying address the needs of different audiences:

- providers
- students (home and international)
- Department for Education (addressing the public accountability issue).

There was recognition that ‘one size would not fit all’ and that dissemination would need to be aligned with user requirements.

As a final observation, there is no doubt that the panel’s confidence in the robustness of its assessments has grown over the two years of the pilots, which emphasises the need for comprehensive training for new panel members if the process is continued and scaled up for all providers.

Social Sciences, and Natural and Built Environment subject panel report

Executive summary

Introduction

The Social Sciences, and Natural and Built Environment (SSNBE) panel comprised 13 academics, eight student members and three further members representing employers and PSRBs. The panel met on 26 and 27 March, 30 April and 1 May, and 30 May 2019 to produce subject-level ratings under the piloted model of the TEF. The panel was co-chaired by Professor Neil Ward, Deputy Vice-Chancellor of the University of East Anglia, and Professor Christopher Hughes, Pro Vice-Chancellor Education at the University of Warwick, with Diarmuid Cowan, the former Students' Union President at Heriot-Watt University and Melz Owusu, former Students' Union Education Officer at the University of Leeds, as deputy chairs. The panel assessed 98 submissions across six subject areas:

- Agriculture, Food and Related Studies (11 submissions)
- Geography, Earth and Environmental Studies (12 submissions)
- Architecture, Building and Planning (22 submissions)
- Sociology, Social Policy and Anthropology (23 submissions)
- Economics (14 submissions)
- Politics (16 submissions)

Evidence (metrics and written submissions)

The panel considered the relative merits and utility of different types of evidence available to support their assessments. There was some discussion about the balance of weightings between some of the core metrics; for example, the relative weightings of continuation and employability and teaching quality. Some contextual information (e.g. employment maps) was not widely used. The pilot provided an opportunity to gain further insight into what types of evidence were most compelling within written submissions.

The assessment process

There was a general level of confidence in the assessment process across the majority of panel members, although this was not universal. There were some concerns about the scope for variability in what assessors were looking for, and a more systematic approach for assessing the narratives of submissions was suggested. The more limited role of employer and professional, statutory and regulatory body (PSRB) representatives was also discussed.

Ratings and statements of findings

The panel felt that the three ratings (Gold, Silver and Bronze) system was simplistic and that greater differentiation, perhaps through a grade-point system or at least through a larger number of ratings, could be beneficial to the assessment process and more helpful to users of the TEF. Ratings for the three aspects of quality were straightforward to produce and were considered to be a positive and helpful development. Statements of findings (SOFs) were felt to be time consuming and there were concerns about their clarity of purpose, utility and scalability. SOFs should also

provide a useful resource for providers and subject areas to inform their approach to written submissions in future.

Potential impact

The panel members considered that subject-level TEF would be an important stimulus to enhancement of education within providers and would also help focus attention on differential experiences and outcomes among different types of students. However, there were concerns about potential unanticipated consequences of the TEF system, especially with the current rating system, including the risk of providers winding down provision in their lowest rated subject areas rather than seeking to enhance provision.

Key findings

For future consideration, the panel suggested the following:

- The balance of weightings between different sets of metrics: it was suggested that continuation and employment metrics could count for relatively less than they did in the subject pilot, and teaching quality could count for more.
- Information on attainment gaps in firsts and 2:1s between different types of student groups could be included within subject metric workbooks.
- Consideration of how student partnership and ensuring positive outcomes for all can be more strongly incorporated throughout the three aspects of quality and the 11 criteria.
- The rating system of Gold, Silver and Bronze could be replaced with either a larger number of ratings or a point score system to greater incentivise enhancement within providers and subject areas. Ratings for the three aspects of quality should be retained.
- The system of SOFs could be revised to reduce the burden on panel members.

Evidence, including metrics and written evidence

In its reflections on the pilot, the panel considered the types of evidence used in the assessment process. This included the nine core metrics, the supplementary metrics and contextual information provided within the metrics workbooks, and the evidence marshalled by providers within their written submissions.

Core metrics

The panel members did not feel ‘overloaded’ with metrics and felt there was value in the spread of the metrics across the three aspects of quality. Some key considerations were as follows:

- **Small cohorts:** Where small numbers of students were included in the metrics, the validity of the data felt much more questionable. The cases for which the panel was unable to arrive at a rating were typically those with the smallest numbers of students in the metrics.

- **Continuation:** There was concern that the double weighting of the continuation metric gave too much weighting to continuation compared to other metrics, particularly in determining the initial hypothesis.

- **Employment:** It was felt that the three core metrics on employment/employability gave too much weighting to this particular aspect of quality. Notably, an employer representative on the panel felt strongly that too much weighting was given to these three metrics. A reduction in the total weighting of the employment metrics need not mean less emphasis on embedding professional skills, as these should be an integral part of TQ and LE. There was some concern about the long data lag around Longitudinal Educational Outcomes (LEO) data as it relates to graduates who studied on programmes several years ago and is a dataset that lacks regional benchmarking. It was felt that the actions of individual providers and subject areas within providers may have only limited influence upon the sustained employment or average earnings of graduates over a prolonged period after graduation.

- **Teaching quality:** Panel members questioned whether the TQ metrics were sufficiently weighted. Teaching quality (TQ1-5) formed a significant proportion of the assessment criteria (5 out of 11) and tended to be covered more extensively in written submissions.

- **Provider-level and subject-level influence:** Panel members felt that learning resources were potentially under less control of subject areas and more under the control of the provider as a whole. Panel members were concerned that subject areas were being assessed on factors which they may have only limited influence over. There was also some discussion about the degree to which student voice and student partnership was shaped at institutional level or subject level.

- **Benchmarking:** There was widespread concern among panel members that the precise basis for benchmarking of different core metrics was not widely understood by panel members, nor was the way that the benchmarking factors differed between different metrics (as set out in pages 23-24 of the 'TEF Subject-level pilot guide'[7]). This risked core metrics and flags being taken at superficial face value without a clear enough understanding of how the benchmarks had been arrived at.

- **Z-scores:** The 1.65 Z-score indicator was considered to be useful, and the yellow shading in the split metrics was especially clear and helpful.

- **Understanding subject areas:** There was some concern that students on joint honours or combined honours-type programmes were apportioned across different subject areas, appearing in multiple workbooks. It became difficult for assessors to interpret data and conceptualise the subject area when it was effectively made up of 'parts of students' rather than whole students.

Supplementary metrics

There was some discussion of the fact that information on attainment gaps in degree outcomes (firsts and 2:1s) was available at provider level for the provider-level panel, but this information was

[7] Available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-subject-level-pilot-guide/.

not visible to subject panel assessors. While gaps in student satisfaction, continuation and employability between different categories of students were discernible through the split metrics data, 'good' degrees were not. It was felt that this would be useful information.

Contextual information

Panel members often found the employment maps difficult to interpret. Only limited use was made by panel members of the maps, although some panel members did find them useful in some particular circumstances.

The panel felt it could be helpful if summary guidance was provided on the key types of accreditation that were significant in some subject areas (e.g. Architecture, Building and Planning). PSRB representatives may be able to help in the production of brief contextual summaries which could include information on what is expected in the subject (e.g. standard PSRBs, learning resources, typical teaching activities).

Evidence in written submissions

Written submissions were generally straightforward to navigate and provided a welcome complement to the metrics and an opportunity for individual provider circumstances to be explained.

It was difficult to identify where the voices of students were clearly coming through in the written submissions. It was common for submissions to include descriptions of the work of Student-Staff Liaison Committees, for example, but compelling examples of impactful student partnership work were relatively rare. Quotes from individual students, or external examiners, were generally not considered substantial enough in and of themselves.

There was considerable discussion among the panel of whether student partnership and positive outcomes for all should be considerations that run through all three aspects of quality (TQ; LE; SO) or whether they should be stand-alone aspects of quality in themselves. Views were expressed on both sides, but there was a strong consensus among panel members that in developing the TEF evidence and assessment process, greater weighting should be given to these two factors.

Where there was non-reportable data, subject areas often struggled to adequately utilise data of their own. Panel members felt that more detailed guidance would help smaller providers and subject areas complete submissions when they had non-reportable data. This could include guidance on how to qualify the robustness of alternative data (where population sizes would need to be clear, for example).

The panel also considered the treatment of criteria for which there seemed less metricised data. These included, for example, valuing teaching and rigour and stretch. When numbers of Higher Education Academy Teaching Fellows were reported, for example, it was sometimes unclear what proportion of staff held such a qualification.

The assessment process

The panel operated well and there was a range of expertise and well-informed discussion to arrive at ratings. In reflecting on the assessment process, the panel agreed that the handling of positive outcomes and differential satisfaction among different groups of students could be a stronger

consideration throughout the assessment process and ideally could be considered across all 11 criteria.

At the outset of the assessment process, panels need to be clear about what the features of submissions are that are important. The TEF subject-level pilot included detailed guidance for assessors across all subject panels, and it was recognised that there needed to be consistency in assessment across these panels. However, an explicit discussion around what is important and how a panel should approach the assessment would be valuable at the start of the process and potentially reduce the risk of metrics capture and make approaching the written submissions more systematic.

It was noted that because employer and PSRB representatives were not part of the initial assessment, they sometimes felt relatively less involved in the assessment process. The role of these specialists in the assessment process would warrant further consideration by the OfS.

Panel members felt that the system of assessing in groups of three and then nine worked well. However, when ratings were finalised by the full subject panel, there was a concern among some panel members that outcomes felt less robust. There was a worry here that fatigue could set in for panel members. Panel members struggled when looking across numerous cases they had not previously assessed and felt less confident to comment on them.

The panel adopted a different approach for agreeing ratings between the first and second batches of assessments. In first batch of assessment, ratings were more likely to be finalised at the second stage of the assessment process, which left the final stage feeling more formulaic. Therefore, for the second batch, the panel deliberately kept a larger proportion of submissions open to further assessment before finalising a rating, and this meant more meaningful work and more richly informed decision-making took place at stage 3.

Although mindful of the data limitations and the constraints of a five-page written submission, panel members generally felt comfortable with the structure and nature of the assessment process and the ability to produce robust ratings. One panel member, a deputy chair, expressed more fundamental concerns about the confidence they could place in the assessment process given the subjective positions and perspectives that assessors brought with them.

Panel members felt that the treatment of widening participation (WP) could be strengthened in the assessment process. For example, there could be an explicit step in the assessment process which is focused on WP. Panel members felt it would also be helpful if the aspects of quality were re-ordered to give greater prominence to WP and positive outcomes for all, which would preferably come first and not last.

Some panel members raised concerns about coming to a single judgement when there were significant numbers of part-time students alongside full-time students. They felt there may be an impact of having a large proportion of part-time students, who may have materially different experiences to their full-time peers. There was evidence of some institutions not focusing sufficiently on these students. Similarly, for subjects that combine multiple areas of provision, such as Architecture, some submissions focused very heavily on the Architecture students but with very little written about the other subjects like Construction Management or Quantity Surveying.

Some panel members expressed concerns about the composition of the panel. There was a discussion on the impact of potential biases, including unconscious bias, the make-up of the panel

and how to ensure that panels are widely representative of the population. There was recognition of structural issues in the sector (such as few black and minority ethnic professors in the sector). The OfS TEF team should continue to strive to improve the diversity and balance of panels.

Ratings and statements of findings

Overall ratings

Panel members considered the Gold, Silver and Bronze rating system to be too simplistic. The rating system produced unfortunate 'cliff edges' and a wider range of categories or a form of grade-point system should resolve this issue. It was also suggested that Gold, Silver, and Bronze could be replaced with words for the levels of performance they indicate, namely Outstanding, Excellent and Good, so that the third rating loses the negative connotations of Bronze.

It was agreed that some submissions that related to new courses were being submitted for a rating too soon. A subject area should have graduates and at least some employability metrics before it is required to be submitted and assessed. Similarly, if a course is closed and no new students are being admitted, it is advised that it should not be eligible for inclusion in subject-level TEF.

Ratings for aspects of quality

Panel members strongly favoured the new approach to producing ratings for the three aspects of quality. Crucially, the panel arrived at the overall rating first and then produced a rating for each of the three aspects. This was relatively straightforward to do. It did not spark extensive and difficult deliberation. The process was helped by the fact that there were two extra ratings (Gold/Silver and Silver/Bronze). It was generally felt that the ratings for aspects of quality would provide useful and more detailed feedback to providers and subject areas which could help inform enhancement strategies.

Learning Environment (LE) tended to score lower as an aspect of quality than the other two aspects. The panel felt that this could reflect differences between local and central provision within institutions and thus where it becomes harder for subjects to demonstrate what is strong or special about the LE aspect of their provision within their local subject area.

Statements of findings

Producing statements of findings (SOFs) was a time consuming part of the panel's work and panel members were concerned about the robustness of the process, particularly for the first batch of assessment, where assessors were given a freer hand to produce SOFs. It proved difficult to co-ordinate input to SOFs across the three reviewers in the assessing trio, for example. It was also harder for lead reviewers to produce SOFs when they held an outlying view within the three or four assessors that had considered the submission. There was also concern about how to manage the variability in the first batch of SOFs, and the potential risk of challenge by dissatisfied providers.

For the second batch of assessments, SOFs were considered to be formulaic, calling into question why assessors needed to be involved in their production, and how useful they would be to providers.

Panel members questioned the value of SOFs and the lack of clarity about who the key audience for them would be – providers who might use SOFs to help inform enhancement strategies, or

prospective applicants, who might use SOFs to inform their choice of course. Panel members felt that the two potential audiences would have different requirements from SOFs, and it was questioned whether SOFs adequately served either purpose. To inform enhancement, it was felt that the ratings produced for the three aspects of quality were likely to be useful. Overall, there was scepticism about SOFs and concern that they would prove difficult in the scaling up of subject-level TEF.

In producing SOFs, most panel members stated it had been difficult for lead reviewers to engage in the discussion and take notes at the same time. If the approach of the first batch were to be maintained, there should be support or guidance on providing for this.

Potential impact

It was widely agreed that subject-level TEF would provide an important stimulus to enhancement within higher education providers. The existence of ratings across providers' subject areas would allow students to make a more informed choice. Panel members felt that the split metrics by student characteristics are likely to drive greater emphasis on differential experiences and outcomes and stimulate strategies for greater inclusivity in the delivery of higher education.

Panel members expressed concern about potential unanticipated consequences of subject-level TEF, especially when combined with the current rating system. It was questioned whether providers might begin to withdraw educational provision from Bronze-rated subject areas in order to improve their overall institutional TEF profile. There was also discussion about the potential interpretation of 'no rating' and whether this could be wrongly interpreted as 'less than Bronze' when in fact it meant that there was insufficient data to arrive at a reliable rating. 'No rating' was considered difficult to explain to prospective students too.

Student panel members considered that TEF ratings are unlikely to be a significant factor in influencing student choice of course and provider for UK students. However, it was felt that ratings and rankings have more purchase in some international markets, although international student data is not included in some metrics, particularly those covering employment.

Subject-specific considerations

It was noted that there was not as much subject-specific discussion as may have been expected. Architecture and Building/Construction submissions required careful treatment as few panel members were familiar with the distinctive structure of architecture programmes and the implication for the interpretation of metrics in this subject area (especially NSS and continuation metrics). Panel members felt it would be helpful to have contextual information about subject areas, including the role of PSRB accreditations and distinctive approaches to delivery of teaching.