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

Annex A: Main panel chair's report

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

Publication date 21 January 2021

Contents

This report is authored by the TEF subject-level pilot main panel chair, Professor Janice Kay, with input from main panel members.

The main panel comprised 38 members in total:

- Chair and deputy chairs: The panel was chaired by Janice Kay, Provost and Senior Deputy Vice-Chancellor of University of Exeter. Professor Helen Higson, Provost and Deputy Vice-Chancellor of Aston University, and Josh Gulrajani, former Students' Union Vice-President (Education) of University of Essex, were deputy chairs.
- Subject panel chairs and deputy chairs: 10 main panel members were also subject panel chairs (academics) and 10 were deputy chairs (students).
- Main panel members: of whom eight were academics, three were students, two were widening participation experts and two were employment experts.

The panel met on 25-26 April, 20-21 May and 13 June to decide provider-level ratings and to moderate across subject panels under the second pilot model of the Teaching Excellence and Student Outcomes Framework (TEF). The panel assessed 43 provider-level submissions and performed a moderation function across all 45 pilot providers.[1]

This report is an overview of the pilot process from a panel perspective, including main panel feedback, with specific feedback from subject panels, students and experts covered in separate reports.[2] Where applicable, key findings from those reports have been synthesised and included here to inform the chair's findings.

General observations

1. Now we have finished the TEF subject-level pilots, it is important to look back and reflect on the substantial amount of work put in to get us to this point, to consider the challenges we grappled with and the lessons we learned along the way. Through the first year of the pilots we tested two models of assessment with 50 institutions, concluding that neither model was fit for purpose, and that a more comprehensive model should be explored. The second year of the pilots brought new challenges. We wanted to test as many potential refinements as possible which required panel members to get to grips quickly with several new features such as an expanded set of core metrics. Through the pilot, we were particularly keen to examine whether the new model could be scaled up to cope with the considerable number of subjects and institutions that a full subject-level exercise would involve.

2. In principle this model – in which all of a provider's subjects are assessed individually – worked well. Subject panels felt well-equipped to make assessments of subjects, and the subject panel structure and process made it possible to assess subjects in a logical framework. A mixture of academics, students and employer and professional, statutory and regulatory body (PSRB)

[1] Two of the 45 participating providers were single-subject providers so were assessed and given ratings only at subject level under this model.

[2] See Annexes B, C, D and E, available at: www.officeforstudents.org.uk/publications/tef-findings-from-the-second-subject-level-pilot-2018-19/.

representatives brought fresh insight and deep expertise which helped provide the nuance that may otherwise have been lost in an exercise operating on this scale.

3. Despite navigating heavy workloads and an increasingly complex set of information, panel members worked with enthusiasm and professionalism. TEF could not operate without their commitment and I would like to thank them for all their hard work throughout. Everyone needed to hit the ground running this year, and with 80 per cent of panellists having worked on the first pilot there was a sense for them of having to 'unlearn' old ways of doing things while learning new information. New panel members particularly had a steep learning curve, and how we could make this easier and more effective is considered below and in the subject panel reports.

4. Both years of the pilot have been conducted in the spirit of 'action learning' and constructive problem solving and feedback from panel members has been key to making productive changes and identifying issues. The evaluation of the pilot is enhanced by the contributions of panellists throughout the process, which were thoughtful, insightful and perceptive.

Key finding: In principle, the model of assessing each subject individually worked well. The subject panel structure allowed for in-depth assessments by individuals with relevant subject insight and expertise.

Training and calibration

5. Effective training of panel members is crucial to running a successful process, and a new approach was trialled this year with online videos and tests complementing the face-to-face training sessions of previous years. This approach was largely welcomed, but the transition to online and remote training needs careful consideration to ensure it meets the needs of all panellists. Further training on interpreting providers' own qualitative and quantitative evidence in submissions should also be developed as the current crop of materials were weighted heavily towards understanding OfS-produced metrics. The extra support would help panel members feel more confident in making their holistic judgements.

6. The calibration exercises succeeded in providing a consistent starting point for panellists' assessments, and this should be built upon in future as a cornerstone of a robust assessment process. It would be prudent to offer further training for panellists on the necessary skills required to be an effective panellist, including negotiation, engagement and creating an inclusive environment, which would empower all panellists to participate fully from the start. Furthermore, the complexity of the data used in the process means that new panellists will require thorough technical training and support as the Business and Law, and Education and Social Care panel states:

'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.'

7. Further work is required to make the process accessible and open to all. It was acknowledged that not all panels were as diverse and representative of the sector as they might have been. Although the panels generally felt comfortable in undertaking assessments with the current representation in membership, ensuring there are voices from underrepresented parts of the

sector and society would further strengthen panels' ability to make fully informed judgements about excellence in higher education. Recruitment of a diverse pool of students for the future framework should draw on the expertise of current student panel members and involve them in training exercises or mentoring.

8. An improved training offer and clearer guidance about roles and responsibilities should be used to ensure that the TEF process is an exemplar of best practice in inclusivity. As simple illustrative examples, panel members will want to have access to a laptop or other device to access online systems and assessment materials during the panel assessment meetings but it should not be assumed that all panellists will own a laptop or similar device, and it should be made clear at the start of the process what will be required and what technological provisions the OfS will make. The meetings themselves require quick navigation and assimilation of written and numerical information, sometimes at short notice, presenting challenges to panellists with diverse needs and additional support should be put in place. Suggestions for additional training were made in the student report, and should be considered in future:

- Chairperson training
- Equality, diversity and inclusion training, including unconscious bias training
- Self-confidence, resilience and negotiating skills workshops
- Data literacy training

Key finding: A thorough induction programme for future panel members will be crucial to the success of the next exercise, with an expanded package of training to be considered. A more diverse membership within panels will support robust decision-making and an improved training offer will support an inclusive assessment approach for all panel members.

Conducting the assessments

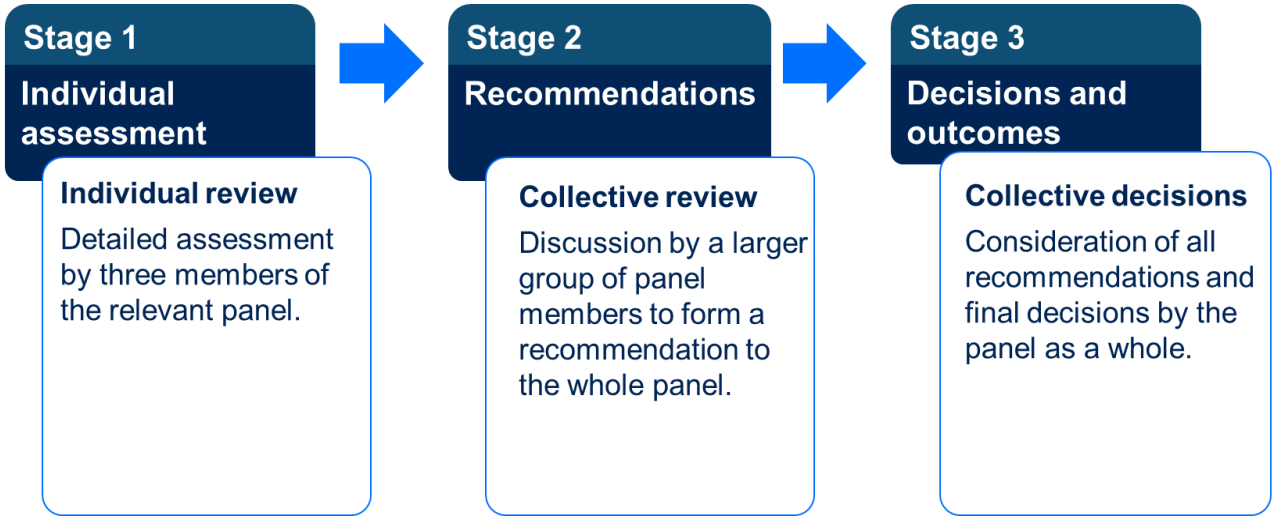
9. The principal role of the main panel was to assess cases at provider-level and to moderate assessment decisions across subject panels. Subject panels were responsible for assessing subject cases within their areas. It is anticipated that ten separate subject panels would be needed for a full subject-level exercise but for the pilot exercise subject panels were paired to create five joint panels (e.g. Arts plus Humanities)[3] to simulate a workload that better reflected what might be required in a full exercise under this model. All the subject panels assessed their cases in two 'batches' in successive months.

10. In all TEF exercises to date, panels have followed allocation and assessment processes that enabled decisions to be honed through a staged collective process (illustrated in Figure 1). In this pilot each subject-level or provider-level case was first assessed by a 'trio' comprised of two academic panel members and one student member (Stage 1). These assessments were

[3] The full panels are listed in Table 5, TEF subject-level pilot guide, available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-subject-level-pilot-guide/.

conducted independently and ahead of the assessment meeting. At the meeting, one of the trio then presented their assessment of the case for consideration within a group of nine[4] panel members (Stage 2), before recommending a rating to the whole panel who would consider and confirm a rating (Stage 3). This process was a condensed version of the TEF Year 2, 3 and 4 provider-level assessment process in which trios met and discussed cases before presenting them to the group of nine. There is great value in this trio discussion as it allows assessors to work through clarifications and test their individual judgements on a case before presenting it to the wider panel. In the pilot, the chair and deputy chair oversaw discussions through Stages 2 and 3, co-ordinated meetings and feedback with the support of TEF officers and OfS personnel, and reported to the main panel for moderation across subject panels.

Figure 1: Stages of assessment



11. Timescales in this pilot were very tight, with almost 700 submissions to assess over a period of three months. In the first batch of meetings, some subject panels found going straight from individual assessments into Stage 2 meeting case discussions with a group of nine somewhat rushed. Subsequent Stage 3 discussions, where the whole panel considered the rating recommendation from the group of nine, meant that some panellists were not involved in a case discussion for significant periods of time. This improved in the second batch of meetings, in which expectations of what level of discussion and input were needed at Stage 2 and Stage 3 of assessment were clarified. However, at both stages, subject panels felt there was a great deal of business to get through in a limited amount of time. Simple cases could be resolved through consensus but, as a consequence of the limited time available, more complex cases were often resolved through compromise.

12. The main panel had fewer time constraints in their provider-level assessment meetings due to a lower volume of assessments to complete. There was opportunity to conduct the face-to-face case discussions between the trio who had assessed a case (as in TEF Year 2, 3 and 4 assessments), before proceeding to Stage 2 with all provider-level assessors, and then recommending ratings to the whole panel at Stage 3. Where the time was available, the three-stage in-meeting process following independent case assessments worked well and allowed for shorter discussions on the clear-cut cases and longer discussions on the more complex

[4] The Arts and Humanities panel had an additional trio so one of their Stage 2 groups had 12 members. Chairs and deputies, plus employers and PSRB panel members, were in addition to the group of 9.

cases. It also allowed employment and widening participation expert panel members to focus on trio discussions where specific issues were identified. The main panel members felt the staged, in-depth discussions led to more robust decisions being made.

Key finding: A three-stage face-to-face assessment process, following individual assessments, supports thorough reviews. Panels need sufficient discussion time to make robust rating decisions, especially with complex cases. Scaling up a three-stage assessment model requires considerable additional panel resource. Further process modelling is required to achieve optimal use of panel members' time.

Subject-level considerations

13. Specific subject-level considerations are explored in detail in each of the subject panels' reports, but a common theme emerging in the second pilot was subject categorisation issues. Although some panels found the new CAH2[5] subject structure (revised for the pilot) helpful, most panels had at least one CAH2 subject where the level of aggregation was challenging, such as Medical Sciences and Nursing and Allied Health (MSNAH) here:

'There was concern that 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.'

14. Interdisciplinary provision was also an area for further exploration this year. The role of the 'interdisciplinary liaison' was introduced and additional data on the level of 'interdisciplinarity' within a subject made available to panels. Interdisciplinary liaisons undertook a deep dive into cases where provision spanned subjects across subject panels. This exercise was interesting but ultimately providers were responsible for illustrating the nature of their provision within a subject submission and this was not always done effectively. It is clearly challenging for providers to articulate provision where their institutional structures or course designs do not align well to the subject structures used in TEF, and this creates further difficulties for panels in assessing such provision. Given the above issues with subject categories, it looks incredibly difficult to fully capture and rate interdisciplinary provision in this model of assessment.

Relationship between provider and subject-level assessments

15. As part of a final 'holistic judgement' step in provider-level assessments, the main panel compared the full set of subject ratings for each provider with the recommended provider-level rating to test their judgement and gauge the degree of consistency between an institution's provider-level and subject-level ratings. In the first year of the subject pilot, the main panel rejected a formulaic provider initial hypothesis derived from the ratings of its subjects. In this second pilot, the main panel judged the profile of subject ratings could influence the provider-

[5] The subjects assessed in the subject-level pilot were based on the HESA Common Aggregation Hierarchy at level 2 (CAH2). In this year's pilot, an amended version of the CAH2 was used, based on feedback from the previous year of pilots.

level judgement. In the majority of cases, however, the profile of subject ratings had little impact on the final provider rating.

16. In most cases there was consistency between provider and subject ratings. Some variation was considered acceptable, and indeed likely to reflect the true situation at the provider. In a number of telling cases, however, the final provider rating was different to all of the provider's subject ratings, resulting in extensive debate in the main panel. Although provider and subject-level assessment made deliberate use of different evidence and criteria, which was welcomed across the main and subject panels, concerns centred on the potential for quite different ratings being awarded at provider and subject-level to undermine credibility of the outcomes for a general audience. Ensuring appropriate and credible coherence between provider and subject-level ratings, particularly for single subject or near-single subject providers, requires further careful consideration.

Key finding: The relationship between provider and subject-level judgements needs to be clarified for all audiences. Large disparities between provider-level and subject-level outcomes risk undermining TEF's credibility.

Ensuring consistency in assessments

17. While the main panel played an important role in moderating provider-level and subject-level assessments, it agreed early on that it would primarily seek to ensure subject panels were assessing and decision-making in a consistent way rather than overriding subject panel decisions. Subject panel chairs and deputy chairs met at main panel meetings to report to the main panel and to review the progress and behaviours of their subject panel. Chairs and deputies found this helpful in considering where their panel might be an outlier or have different thresholds for ratings. Despite this, some differing panel behaviours emerged naturally, and the current moderation processes could not ensure total consistency across all panels. The main panel members felt that, while they were able to operationalise the moderation mechanisms proposed in the pilot, the current approach was insufficient for ensuring an appropriate level of consistency in a full exercise with published outcomes.

18. Within all panels, several cases were assessed twice by different groups, serving as a cross-check for moderation purposes. This was a useful exercise as it allowed panels to identify where their internal boundaries for each rating lay and to check their collective judgements. However, most panels felt that more cross-checking could have been built in at an earlier stage, such as during a more extensive calibration process, and checked again throughout the process, as noted by the Medical Sciences and Nursing and Allied Health panel:

'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.'

19. As an additional measure of consistency, a deep-dive exercise was conducted by a subset of main panel members, who each focused on one provider and its subjects. This was a useful process which helped main panel members understand subject-level issues. However, the

impact of scaling up the moderation process using a deep dive and increased calibration and cross-checking mechanisms would be considerable in terms of planning and resources.

Key finding: The moderation mechanisms tested in this pilot were helpful but insufficient to achieve the consistency of judgements required by a future subject-level TEF. Scalability is impacted by any need for additional moderation mechanisms such as meetings, processes and resources.

Student engagement

20. The role of main panel student deputy chair was introduced this year after student panellist feedback on last year's process. Josh Gulrajani (former Students' Union Vice-President, University of Essex) was appointed to the role. This worked well and reflected the importance of parity of student and academic roles in the exercise. Josh played a key role in chairing assessment meetings, co-ordinating student panel member feedback, and producing a separate student report.[6] Its main findings include the positive impact of increased student voice in all aspects of the process, the need for an improved package of training and information for student panellists, and the need to improve panel diversity.

21. The introduction of the National Student Survey (NSS) student voice metric was welcomed unanimously across the panels. The addition of a new criterion, TQ5 (Student Partnership), highlighted and focused attention on the importance of student partnership. However, while there were improvements in student engagement with provider-level submissions, subject panels were frequently disappointed with the lack of student engagement coming through in subject-level submissions. There were also different views about where 'student voice' should sit within the process, with interesting alternatives considered by the Social Sciences and Natural and Built Environment panel:

'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 (Teaching Quality; Learning Environment; Student Outcomes) or whether they should be stand-alone aspects of quality in themselves.'

22. Another improved element of student engagement in the pilot, with a direct influence on the main panel processes, was the student declaration. Student members had made the declaration, to be completed by student representatives[7] at participating providers, a recommendation from the first pilot. It was designed as a mechanism for students to demonstrate their involvement with the TEF process independently of the provider submission, and all panel members welcomed its introduction to this year's process. As the declarations were made at provider-level only, student main panel members held a session before the first main panel meeting to discuss how the declarations might be used in provider-level

[6] Available at: www.officeforstudents.org.uk/publications/tef-findings-from-the-second-subject-level-pilot-2018-19/.

[7] Participating providers were asked to nominate a student, ideally a representative from their students' union or association, as their lead student representative for the pilot and to complete the student declaration.

assessment, deciding that it should be used by all main panel members to supplement the other provider-level assessment materials.

23. In practice it was found that the student declarations in their current form were not entirely fit for purpose, but that with a redefined scope they could be more informative. While it was helpful for the panel to understand how students had been involved in the TEF process, it would also have been useful to understand how the provider engaged its students in partnership more generally. The panel would have liked to see a student submission that contextualised and commented on students' engagement with teaching and learning at their provider. The student findings report[8] covers some of this main panel discussion:

> 'Panel members felt there were legitimate questions that could be asked relating to the students' involvement more broadly in the enhancement of teaching and learning. The student declaration could not accurately be used as a proxy for this, as it was not created for this purpose.'

24. An additional consideration on student declarations is that the panel did not decide on a firm approach to cases where participating providers were unable to identify a student representative, or the student representative opted not to submit a declaration. In the pilot, the lack of a student declaration could not negatively affect a provider's final rating. However, in a future exercise where a student submission could be used as evidence to inform a judgement, parity issues arise when evidence from students is missing. This must be carefully considered if a student submission is to be developed and recommended.

Key finding: The student engagement measures used in this pilot should be maintained and strengthened in future TEF. How to improve the scope of student engagement in future TEF requires further exploration.

Data, metrics and evidence

25. The information available to panel members increased in both volume and complexity between the first and the second pilot. The metrics workbook contained a substantial amount of information, from the provider contextual data, new basket of nine core metrics and associated split categories, to new attainment data at provider level and course contextual data at subject level. Panel members recognised that the information available could provide rich and valuable insights, but that it could also be overwhelming and difficult to navigate in real time, as in panel meetings or when asked to be an 'extra reader' on a tricky case. Considerable training was needed to get to a place where most panel members were comfortable with the range of information they were being asked to use, and it did not always appear that more information led to more nuanced decision-making.

26. Addressing the volume of information was particularly evident at provider level where submissions were up to 15 pages (compared with a 5-page maximum for individual subjects) and possibly contributed to the slight disengagement of panel members at Stage 2, mentioned

[8] Available at www.officeforstudents.org.uk/publications/tef-findings-from-the-second-subject-level-pilot-2018-19/.

in paragraph 11. However, the 15-page submission allowed providers space to build a compelling narrative with strong evidence (compared with the limits of a 5-page submission), which then enabled main panel members to feel somewhat more confident in the final ratings they awarded.

27. One of the difficulties with an enlarged basket of metrics was understanding which cohorts and which programmes contributed to each metric and how well the provider evidence then mapped onto the students represented in the metrics and the students covered in the contextual data. This could be addressed with further consideration of how to improve the metrics and contextual data presentation and training for both panel members and providers; however, consideration should be given to what information could be removed to ensure panel members can focus on what is most important. For example, some elements of the contextual data, such as the maps, could be removed with little detriment to the core process and would reduce the data overload experienced by panel members.

28. The relative weighting of metrics was also discussed by all panels, with views differing on the weighting of employment and NSS-based metrics, but all panels were concerned about the 'double weighting' of the continuation metric. The new method of generating the initial hypothesis for the basket of nine metrics was also thought by many panel members to make it more difficult to arrive at a Gold final rating: one negative flag in any of the nine core metrics would almost always prevent a Gold initial hypothesis, which was felt to be an unnecessarily harsh starting position.

29. Although panels moved subjects up to a Gold final rating from an initial hypothesis as low as Silver/Bronze in a few cases, the anchoring effect of the initial hypothesis was strong. In most cases, very few subjects with a Bronze initial hypothesis moved upwards. An additional difficulty with the initial hypothesis was that some providers who had substantial numbers of students in both full-time and part-time modes of study had very different metrics for each mode. Panel members found this hard to resolve within the current assessment process with a single rating awarded at the end.

Key finding: More information does not necessarily lead to more nuanced rating decisions. If an evidence source is contributing complexity without adding value, it could be removed. The relative weighting of metrics and the construction and use of the initial hypothesis need further consideration.

Supplementary metrics

30. Two kinds of supplementary data were provided this year. The first was evidence about changes in degree classes[9] over 10, three, two and one years (grade inflation). The second was information about differential degree attainment split by different participation categories (e.g. Mature, BAME, IMD). These were available at provider level only. Inclusion of the differential degree attainment data was particularly welcomed by the main panel, although the panel reflected that supplementary information had seldom contributed to provider-level judgements. The panel felt that the attainment data was difficult to operationalise, as the guidance was less prescriptive about how and when to include it in the assessment steps, and

[9] First, 2:1, 2:2 and below, and unclassified degrees.

written submissions often did not address supplementary data fully or effectively. The grade inflation data was seldom used to inform judgements, and most panel members felt it had limited relevance to their assessment of a provider's teaching excellence. However, given the importance of addressing gaps in degree attainment, benchmarked data on differential attainment should be included as a core metric as a way to ensure it is fully considered in assessment.

31. Though the data on differential degree attainment is complementary to the splits by characteristics in the core metrics, panels found it difficult to reconcile some of the information with TEF criterion SO3 (Positive Outcomes for All). There were divergent views on how far it was acceptable to have a gap in attainment between different groups of students but still meet the SO3 criterion in a holistic judgement, given that gaps in attainment exist consistently across the sector. The OfS's access and participation work focuses on gaps for underrepresented groups and does not use benchmarking in its datasets, which may lead to confusing messaging around performance and targets for individual providers when compared with TEF data. The TEF process can support institutions to address gaps through their education provision and engagement with students, but the relationship between TEF and access and participation data must be clarified to send an effective message to the sector.

Key finding: TEF must work closely with the OfS's access and participation function if attainment gaps are to be eliminated in the sector. A more developed metric on differential degree attainment would be welcome. TEF does not appear to be the right place to address grade inflation.

32. The debate around Longitudinal Education Outcomes (LEO) data as a core metric continued this year. Region of employment was the main issue discussed by panels, as the metrics use the national median earnings threshold regarding salary, while there is considerable variation in median earnings by region. The employment experts noted that recent changes to LEO meant that the dataset could now be benchmarked by region. It was also observed across the subject panels that one of the measures, 'above median salary', provided information that is not always a motivating factor for undertaking higher education study, and that the quality or value of a degree could not solely be equated with salary outcome. The main panel employment experts offered a nuanced look at LEO issues, recording in their report that "there are compelling-sounding arguments both for and against mitigating for region". Now that this data is available, it seems appropriate that region should be considered when looking at LEO salary data, but the specific method of doing so should be carefully thought through.

Data limitations

33. Although the earlier point (paragraph 24) was made about the process becoming too data-heavy and complex, the problem of missing data and small numbers at subject level identified in the first year of the subject-level pilot remains a key issue. Some metrics were absent if they were not reportable due to cohort size, NSS boycott activity, or data protection suppressions, reducing evidence available to subject panels. Also, evidence was limited where cohorts were small, provision in the subject had changed significantly across the different years captured by the metrics, or provision was relatively new with limited outcomes data. A particular issue was noted around the non-availability of LEO data for some types of providers, which removed two of the three core outcomes metrics. In some cases this was in addition to a very short or

insubstantial submission for a subject, which provided little or no additional evidence. This relates to the earlier point about the formulation of an initial hypothesis: where metrics are missing, the calculation of the initial hypothesis becomes weighted towards a particular aspect of quality or metric type, requiring very careful consideration of all available sources of evidence for a judgement to be made.

34. There is an inevitable tension between the ability of panel members to be confident in their judgements and not to disadvantage providers who offer smaller or more niche subjects, as statistically speaking there is less confidence in the data where there are small numbers. Increasing the threshold of 'assessability' would penalise some providers that offer courses with small numbers. The panels felt that some providers were able to offer good mitigation by including a strong narrative and extra internal data within the submission, but this was not consistent from case to case. The majority of cases with a paucity of data failed to convincingly fill the gap. Panels found that when evidence was scarce, pieces of information that might otherwise have not had a large influence on the outcome gained prominence, but this was not always a fair approach. Panels judged on a case-by-case basis whether there was enough evidence to give a rating, but with 7 per cent of assessed subjects not being awarded a rating they were concerned that this approach would not be scalable.

Key finding: Addressing limitations in the metrics, especially at subject level, remains a significant challenge and the pilot has not identified any viable potential mitigations. Detailed work should be undertaken to consider what information is essential to the process, what can be refined, and what can be removed, as there remains a tension between lack of data in some ways, and too much in others.

Ratings and statements of findings

35. In the second pilot, statements of findings (SoFs) were given to providers alongside their overall ratings and subject ratings. At subject level, two approaches to writing SoFs were tested: a free narrative, with one rating for the subject as a whole; and a brief prescribed statement, but with each of the three TEF aspects of quality[10] judged separately, as well as a rating for the subject as a whole. The main panel used a hybrid approach to SoFs, with the SoF providing ratings and a narrative for each aspect of quality as well as the overall rating. In all cases, one of the three panel members who had originally assessed the case at Stage 1 wrote the SoF.

36. All panel members felt it useful to give more granular feedback to providers and welcomed the use of ratings and borderline judgements for the aspects (Gold/Silver, Silver/Bronze). Creating a five-point ratings scale for each of the three aspects helped soften the 'cliff edges' that can arise from awarding a single overall rating on a three-point scale. However, presenting a profile of aspect ratings against an overall rating that may not match the aspects also has the potential to be confusing, so this should be carefully considered. A further reflection on the potential issues with rating the aspects of quality came from the Social Sciences and Natural and Built Environment panel, although this was not necessarily borne out in the cross-panel data on aspect ratings:

[10] Teaching Quality (TQ), Learning Environment (LE), and Student Outcomes and Learning Gain (SO).

‘Learning environment 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 learning environment aspect of their provision within their local subject area.’

37. Strong concerns emerged around the scalability of providing a full narrative SoF for every subject. It was not possible in the time available to ensure a consistent quality of SoFs across – and even within – panels. This was particularly problematic at subject level, where providers would be receiving up to 34 SoFs each potentially written by a different author. The Natural Sciences and Engineering and Technology panel report reflected the experience of most panels in saying that:

‘…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.’

38. As captured in the quote, there was a sense that the current format of SoFs would not provide effective feedback for enhancement, and that this is potentially an opportunity lost if TEF is to be used for institutional enhancement.

Key finding: Neither approach to statements of findings piloted was proportionate. A scalable solution must be found for providing feedback that is useful to providers, consistent across panels, but not excessively burdensome to produce.

39. Provider-level TEF judgements over the past four years have shown that excellence is found across the sector, with a diverse range of providers achieving Gold awards. As the main panel met to reflect on the ratings awarded at the end of the second subject-level pilot, we considered that larger providers and typically universities were more likely to achieve a higher final rating than smaller providers within this sample. There may be legitimate reasons for this observation, such as lack of regionally benchmarked LEO data that may have affected different types of provider in different ways. Equally, while the sample of providers participating in the pilot was intended to capture the diversity of the sector, and panels felt that they were fair to all types of provider in their assessment, it was nonetheless a small subset and not representative in numbers across provider categories.

Key finding: Further consideration is needed on how the process can be fair and consistent for the diversity of providers.

Enhancement and sector reflections

40. A key driver of the exercise is to recognise and reward excellent teaching, and it was a pleasure and a privilege for all panels when they came across a submission setting out unquestionably outstanding provision. Those providers receiving a Gold rating typically demonstrated characteristics such as:

    a. authentic and progressive student engagement beyond the TEF process

    b. a sense of strategic direction and purpose reflected in both the data and the narrative

    c. proactive, transparent and honest responses to areas of lesser performance

    d. deep knowledge of the students studying at the provider

    e. unequivocal demonstration of impact to ensure the very best student experience in their own context.

41. The Business and Law and Education and Social Care panel articulated the wider feeling that providers should feel confident in articulating both what went well, and what hadn't 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'.'

42. Panel members observed that, in some cases where providers and their subjects were close to Gold, but not quite there, it was too early for real impact to be shown, or impact was not fully articulated. Panellists also observed many examples of standard good practice, innovation and enhancement. Nonetheless, this did not always translate into better submissions, with several panels noting that across the two pilots it was clear that some providers struggled to articulate their strengths and weaknesses. For this reason, panel members would have liked a mechanism to provide focused, private feedback on less strong submissions.

43. Given the time pressures in the pilot process, it was found to be especially hard to capture stand-out elements of good practice or enhancement to reflect in SoFs under the current model, as summarised by the Arts and Humanities panel:

> '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 benefitted from an OfS officer or panel member who was responsible for recording examples of best practice to be compiled and shared.'

Key finding: The process should be amended, or new panel roles introduced, to allow the systematic capture of best practice and enhancement information.

Conclusion

44. The pilots have been a fascinating experience. Even in piloting we have already uncovered exciting examples of excellence at subject level, as well as helping institutions identify areas for improvement. Excellence can clearly be found throughout the sector and providing information at subject level would add value to provider-level awards. Testing out processes and the validity of the models has been vital, given that considerable issues with subject-level assessment have emerged across the two years.

45. It is essential that the exercise remains proportionate and timescales and resource have been an ever-present concern, particularly when considering the full volume of subjects and information that would need to be assessed if rolling out the current model to the whole sector. There are serious issues with data availability and reliability at subject level which we were unable to solve within the scope of the pilot but must be addressed or mitigated in future if the exercise is to remain robust.

46. The consistency of assessment and moderation across the panels emerged as a complex issue in this comprehensive subject-level assessment model. Additional moderation processes would need to be in place for a full-scale exercise under this model, as while the processes tested in the pilot led to broadly consistent and coherent outcomes, scaling up would lead to more variability, presenting a real risk to credibility.

47. The relationship between provider and subject-level assessment was also re-tested in this model, with no formulaic link between the two levels. Panel members were required to establish and justify the relationship themselves but were unable to satisfactorily resolve the relationship in all cases. The purpose of providing information at the two levels must be clarified for panels to fully understand the connection and feel confident in awarding ratings at both levels.

48. At this critical stage in the development of TEF, these are substantive issues that remain unresolved. They must be overcome or otherwise mitigated in order for TEF to evolve into a robust and respected exercise that holds the confidence of providers and students.