Annex B: Panel reports

Contents

Arts Panel report

Executive summary

1. The purpose of this report is to present the findings of Teaching Excellence and Student Outcomes Framework (TEF) Pilot Arts Panel. Amongst other things, it focuses on the particularities of the arts subject classification. The breadth of disciplines in creative arts and design is considered and the lack of differentiation between arts and creative arts and design is questioned. The implications of separating performing arts into a new Higher Education Classification of Subjects (HECoS) Common Aggregation Hierarchy (CAH) category are discussed.

2. The Arts Panel concludes that Model B, with adaptations, has the most potential to be fit for purpose since it offers the most rigorous approach to higher education provider subject ratings.

3. This panel had its own widening participation (WP) expert as a core member who was highly valued and it is recommended that this role is replicated on all subject panels.

4. The panel makes a range of recommendations for the TEF criteria and grade descriptors to improve their fitness for purpose in the context of subject-level TEF.

5. There are several points in the report where the emphasis of the student perspective is highlighted to ensure that stakeholder views are fully recognised. The students were particularly interested in why some submissions were not ‘student friendly’ and they noted that several submissions gave little indication of what it felt like to be a student in that provider.

The role of metrics in subject-level TEF

Eligibility for rating

6. The adoption of 35 Common Aggregation Hierarchy 2 (CAH2) subjects is unsurprisingly associated with issues of low levels of reportable data and suppressed metric workbooks. There was an inverse relationship between the size of student cohort and the panel’s confidence in the rating. Lowered levels of confidence associated with small cohorts lead this panel to propose that there is a student cohort size threshold below which a subject is not eligible for a rating. The panel recommends that this threshold is set at 100 but acknowledges that this will be challenging for further education providers.

Non-reportable metrics

7. The panel expressed concern that giving a rating to subjects with non-reportable metrics lacked rigour and would not help student choice. In these cases it was assumed that subjects below threshold would then carry the provider rating but there was concern that this would lead to ratings of Silver in subjects that might actually be Bronze (we referred to this as ‘Bronze wearing a Silver jacket’). It is proposed that such subjects are given a TEF award of ‘No rating’ but student panel members were concerned about what a prospective student might infer from a course that has no subject rating.

Courses that are not recruiting

8. The panel noted issues in relation to courses 'running out'. The rating a non-recruiting course is given is of little value to prospective students if this rating is then applied to a new course in this CAH2 area. In one case the panel were not able to reach a rating for this reason.

Absence and inference

9. It was interesting to note the different ways that panel members (across and within panels) responded to metric absence. One approach observed was to respond to limited data by taking holistic judgement to Bronze (so absence of data lowers rating). However there were panel members who argued that limited data did not equate to poor data and that a quality inference of this kind could disadvantage small but high-quality providers. Absence is not indicative of low quality – it is simply indicative of low numbers.

Metrics and student surveys

10. Some members of the panel were concerned that five out of six core metrics rely on students to respond by filling in surveys and if the students do not respond a small provider has to manage with non-reportable metrics. The continuation data is the only metric that the provider establishes without the need to ask for students to fill in surveys.

Weighting of metrics

11. There was interest in the move to half-weight metrics based on the National Student Survey (NSS) and concern was expressed about the subsequent 'double-weighting' of the Destinations of Leavers from Higher Education (DLHE). This was discussed at several points across the pilot meetings particularly when Teaching Quality (TQ) metrics and Student Outcomes and Learning Gain (SO) metrics reported opposing indicators.

Statistical validity

12. The panel was interested to establish agreement about the statistically valid limits from which a change of hypothesis can be sustained as a result of split metrics variation (both positive and negative).

Graduate outcomes and LEO

13. Whilst the arts higher education community has argued for some years that six months is not long enough to establish an arts based career[1] there was a concern about the use of Longitudinal Education Outcomes (LEO) data. This metric does not include enterprise-based self-employment (which is a common graduate outcome for significant numbers of arts students) so LEO offers a very partial picture of employment. Interestingly there was also debate about what the panel could deduce from the LEO data, given that the cohort whose data was used graduated in 2011. This does not tell us about the provider in the present.

[1] Ball L, Pollard E, Stanley N (Jan 2010), 'Creative Graduates Creative Futures,' https://www.employment-studies.co.uk/creative-graduates-creative-futures.

14. The Standard Occupational Classification (SOC) codes 1 to 3 in arts are improving but there is still a problem that relates to the ways that graduates describe their self-employment and start-up enterprise outcomes to DLHE.

Scope of the metrics

15. It was noted that it is unhelpful to employ metrics that are not available to the full sector. There was a preference for future iterations of TEF to employ metrics that all providers can present.

Teaching intensity

16. Teaching intensity was considered to be of no value. There was only one instance where the Teaching intensity data served to support a submission that talked about the subject's emphasis on work-based learning. There was considerable opposition to a metric that is input rather than output-focused.

Grade inflation

17. 'Grade inflation' is an unhelpful phrase. TEF is premised on developing teaching excellence but the inclusion of grade inflation does not appear to want this development of excellence to enhance attainment. The metric (and expected commentary) is premised on the idea that improved attainment is a problem. All classification data has been agreed through external examiner scrutiny. The response to the grade inflation metrics opened up useful debate about the need for TEF to look at attainment. There was interest in TEF exploring the differential attainment for particular student groups. This would encourage providers to discuss their work to reduce differentials which would be relevant to support 'positive outcomes for all'.

18. Furthermore there was a concern that a punitive focus has the potential to differentially impact on black and minority ethnic (BME) students. The sector has a 20 per cent attainment differential and urgently needs to address this in the coming years. If providers are worried about 'grade inflation' metrics they might delay addressing these differentials. BME attainment is depressed and addressing this will lead to increased attainment. It is imperative that there are no perverse consequences to measuring grade inflation.

Provider submissions

19. The recent introduction of TEF in the higher education landscape means that there is no agreed approach in the sector about who authors the submission and how TEF preparation is managed. This means that the resources and staff time associated with TEF preparation vary enormously. This was particularly evident in the pilot. As a result, in some submissions there was a very poor understanding of what constitutes evidence. The pilot surfaced concerns about the quality of cited evidence from some (but notably not all) of the further education providers.

20. The panel was presented with one case where there was no submission at all and several cases where the submission was so limited or unusual that there was discussion about its status and usefulness in relation to the holistic judgment. There were several cases where the submission was of such limited utility that the panel discussed invoking paragraph 7.65 of the TEF specification.[2] This led several panel members to request that the written submission

[2] Paragraph 7.65 of the Teaching Excellence and Student Outcomes Framework Specification gives a set of rules for reaching a judgement where a submission contains no substantive additional evidence against the

becomes a mandatory part of a TEF submission. The student panel members expressed concern that submission quality suggested that guidance was not clear enough for providers, particularly in relation to what counted as impact in relation to the criteria.

21. The usefulness of student and external examiner quotations was questioned by the panel and it was agreed that individual quotes only constituted evidence when they backed up a wider point articulated with stronger evidence.

22. The ways that providers discussed industry varied considerably and the panel agreed that this was least successful when the submission 'name-dropped' prestigious employers – rather than discussing the demonstrable value of industry links and the benefits for most of the students.

23. The panel reflected on the agency of the authors for subject submissions to change and direct teaching excellence. To what extent are subject leads in a position to lead impactful enhancement against all three aspects of quality at subject-level? This may provide additional justification for reviewing the TEF criteria to ensure they all work at subject-level.

24. The panel also expressed concern when the submissions listed extracurricular offer to students without additional commentary that evaluated take up and impact on learning. Text that simply sets out offer reads more like a prospectus than TEF submission evidence.

25. The panel noted that submissions that referred to provider evaluation (such as internal student surveys) and included response rates were stronger than those which made no reference to response rate. The panel agreed it was hard to evaluate weight of evidence without response rate.

26. Subject-level TEF is particularly challenging for very small providers, but it is also challenging for larger providers where the subject category includes many course titles spread across a number of higher education structures. The panel found it particularly difficult when providers did not list their courses, which decontextualised the submission. The panel recommend that the list of courses in each subject is provided as part of the contextual data.

27. A small number of panel members who were not part of TEF Year Two were uncomfortable about what they regarded as the lack of focus on proving the veracity of the submission and would have preferred to have been offered evidence that was triangulated. This was particularly true for providers that were awarded Gold. This was not the view of the majority of the panel.

The relationship between the metrics and the written submission

28. The panel debated at length the circumstances under which a submission changes the outcome of the rating given to a provider. Some of this debate covers ground explored in TEF Year Two regarding the challenges associated with assessing poorly written submissions and the extent to which a very strong submission serves to mitigate poor metrics. This emphasises the importance of time for deliberation and this needs to be factored in when TEF Year Five

---

TEF criteria (https://www.gov.uk/government/publications/teaching-excellence-and-student-outcomes-framework-specification).

scales up. If scaling up results in there being no time to discuss points such as these then TEF is in danger of becoming an overly metric-driven rating approach.

29. There was a lot of discussion about the circumstances that might result in a rating being lower than the metric-based initial hypothesis. There were two positions. One view was that if the submissions could move subjects up a grade then it was clear that they should also have the potential to move a subject down. This was contrasted with another perspective that felt the submission was an opportunity for the subject to add additional evidence and if this was weak then much of it was discounted and the initial metrics held. The tension between these two views increased when the initial hypothesis was felt to be a 'Silver by default'.

Comparing Model A and Model B

30. The key point to make here is that in CAH2 arts there was minimal difference between the ways Model A and Model B worked at subject panel level. There was the same burden of work for the Arts Panel in relation to both models. In Model B there were two providers that 'ported in' subjects and these were dealt with effectively but where courses were not listed in the submission it did lead to uncertainty about which sections referred to all subjects and which were pertinent only to one of the subjects.

Model A

31. The Arts Panel concluded that Model A was less robust than Model B. The panel noted the challenge of reviewing exceptions when there was no information available about why the subject was an exception. This was exacerbated by the fact that the pilot included some 'test case exceptions'. In this model there was considerable tension between the artificiality of the split between the work of the Main Panel and the subject panel. The Arts Panel acknowledged that knowing the reason for the exception might 'lead' judgement but it did point to the weakness of the model. The panel was keen to see the provider and arts submission and metrics for all providers. Model A uses metrics to arrive at an exception list, a point that the panel was uncomfortable with.

Model B

32. When comparing both models there was agreement by the panel that Model B was more robust. There was consensus that aspects of Model B need to be retained in spite of concerns about burden and scalability. It was noted that there was most consensus about rating profiles in Model B.

33. The focus on the criteria TQ2 (Valuing Teaching), LE1 (Resources) and SO3 (Positive Outcomes for All) worked for the provider statement in Model B and it provided the reader with a useful frame to approach the submission. This suggests that further work to differentiate criteria would be useful.

34. What the pilot has not clarified is the precise nature of the relationship between the parts (a provider's subjects) and the whole (the provider). This is an area that needs further development. To illustrate this there was a query about why there was the need to look at provider metrics given that this already happens at subject-level. However, the use of the subject-based initial hypothesis (SBIH) and the provider metrics and submission scrutiny did work as an approach to arrive at a provider rating.

Model X1

35. The panel sought to identify how aspects of Model A and B can be combined to inform future models ('Model X'). Some panel members suggested that larger interdisciplinary panels might be established that have two subject experts for each subject group. This approach could usefully combine elements of TEF Year Two and Subject TEF. The role of the subject experts would be to moderate and safeguard disciplinary integrity in ratings applied.

Model X2

36. A second, more popular, option is a version of Model B for all subjects with no grouped subject submissions.

Training, preparation and judgement

37. The three-step process was considered to be useful but there was a view that the TEF training focused exclusively on metrics which meant that whilst the panel was confident with the structure and formula that guided its decision making for step 1a and step 1b, some of the panel members would have valued more specific guidance to help them evaluate the submission. Some panel members requested greater focus on the precise nature of the relationship between the metrics, submission and the criteria and would have appreciated clearer steer about the importance of not bringing extraneous knowledge about a provider into rating process.

38. As the panel worked together it developed an approach to verbally reporting the case for each provider rating that stayed firmly anchored in the aspects of quality and the criteria. This enhanced the rigour and transparency in relation to the ways that the submission was discussed and how it contributed to the final rating, particularly when the holistic rating changed from initial hypothesis.

39. Due to the size of the Arts Panel sample there was time for a dialogue to build consensus. It would be interesting to explore the impact of time given to discussion on confidence to move rating from initial metric hypothesis. The question is this: is there a correlation between time for discussion and moving ratings from initial hypothesis as part of holistic judgement?

40. It was noted that some panel members used the borderline rating (as opposed to the confidence rating) to communicate a lack of confidence with the solidness of rating that arose due to lack of evidence (as opposed to a borderline reached in individual assessment in response to a full set of metrics and a well written but borderline submission).

41. The panel suggested that providers should be asked to explicitly refer to splits in the guidance.

42. Some of the panel wanted the providers to be offered a template for the submission. This was a contested point. It may be useful to offer the option of a template that sets out evidence in relation to the three step process and aspects of quality but to stress that this is not obligatory. Beech's (2017) research[3] underlines that TEF Year Two submissions that resulted in a move

[3] See Going for Gold: Lessons from the TEF Provider Submissions, https://www.hepi.ac.uk/2017/10/19/going-gold-lessons-tef-provider-submissions/.

from initial hypothesis to a higher rating were very diverse so it is important that any template offered does not constrain providers or prevent them from setting out excellence on their own terms.

43. We noted the differences between workload across the panels and are interested in the variety of deliberative processes these differential workloads necessitated. TEF Year Two panel members and assessors would like to see a version of the 'ninesome' approach and structure[4] developed and applied to subject-level TEF because it offered a consistency of approach.

44. We also noted that the numbers of assessors looking at each provider varied across panels and again we are keen to see what can be learned from these differences.

45. At different points in the process the panel was visited by various colleagues from the Office for Students (OfS) and the Main Panel. This was useful and could perhaps be further developed to support cross-subject moderation.

Grade descriptors and criteria

46. The panel proposes that the best-fit grade descriptors are adapted for the purposes of subject-level TEF. There was a lot of discussion about whether or not there were non-negotiable sentences in the best-fit descriptor. This point primarily applied to Gold rating, with the key question being: should elements of best fit be non-negotiable for Gold? The panel noted that the grade descriptors could better reflect the non-metric aspects of the criteria.

47. The panel noted that there was a bigger cliff edge between Bronze and Silver best-fit descriptors than there was between Silver and Gold. It is interesting to reflect on whether Silver or Bronze reflect the threshold performance, given that Silver is premised on metrics that are on their benchmark. In the Bronze descriptor it seems anomalous to refer to 'good' and 'below benchmark' in one section.

48. The panel noted that the term 'students from all backgrounds' did not recognise that some providers have a less diverse student population, so this sentence is easier to meet for those providers. It was noted that some providers and subjects are awarded Gold even when there are negative flags in the split metrics. The reference to contact time should be removed from the descriptors and there should be reference to inspirational teaching in the Gold descriptor.

49. Several TEF Year Two panel members have commented that subject-level TEF submission scrutiny did not feel that different to TEF Year Two provider scrutiny, which is surprising and appears to point to the need for criteria that are adapted for purpose. One provider wrote a very strong and coherent articulation of subject pedagogy and the panel noted that no existing criteria directly addressed this. The relationship between generic provider-level criteria and subject-focused criteria needs further development.

[4] The 'ninesome' approach refers to stage 2 of the current provider-level TEF assessment process, where panel members and assessors come together in groups of three and then nine to discuss and refine individual assessment of cases. Recommendations from the groups are then put forward to the TEF Panel in stage 3, where the panel makes final decisions on outcomes.

CAH2 creative arts and design

50. Arts is the only subject group that only has one CAH2 subject linked to it. This means that there is no differentiation between subject group (arts) and subject (creative arts and design). The subject unit of creative arts and design covers all aspects of visual arts, design, performing arts, photography, film, TV production, music and dance. The panel liked the way that CAH2 brought together studio-based teaching and learning practices. In terms of student share of higher education population this is the largest subject (much larger than Celtic studies for example). This means that Model A and Model B play out in ways that are different to those found in subject groups that have between two and seven subjects within their grouping.

51. The key advantage of the current CAH2 approach to arts is that by drawing together these diverse disciplines into one subject the TEF panel is more likely to have reportable data. This increases the robustness of the decision-making and ensures that TEF is rooted in student outcome data. The key disadvantage of CAH2 for arts is that it brings together and merges delivery from across conservatoire, specialist and mainstream further education and higher education art, design and drama. These disciplines have very different approaches to teaching and contact hours so the usefulness of teaching intensity metrics will be very limited. The challenge associated with the breadth of CAH2 creative arts and design surfaced where drama and art and design were located within one provider and it appeared to be the case that what was being reported on was two quite distinct subjects. This also pointed to a concern that arts' breadth and size had the potential to allow some of its sub-disciplines to hide within larger metric workbooks.

52. Whilst the panel concluded that it was able to secure ratings across creative arts and design, if changes are going to be made to CAH2 the panel proposes that performing arts, music and dance are disaggregated into a new subject area. Additionally the panel proposes that architecture and communication and media are brought into the Arts Panel.

53. The panel was concerned about how new and emergent subjects that usefully connect computing and arts (for example, creative computing and creative coding) were poorly served by CAH2. They wanted these developments to be recognised on their own terms – rather than being categorised as computing or arts.

The Arts Panel

54. The panel expertise was very strong and there was mutual respect for expertise and stakeholder perspective. All panel stakeholders had the same workload and this worked well. The allocation meant that everyone played a key role and no-one was simply there in an advisory capacity. Whilst we only had one further education and one alternative provider of higher education colleague on the panel, their contribution ensured that the needs of these providers were understood.

55. The panel self-appointed a widening participation (WP) expert and we strongly recommend that this is obligatory for all subject panels given the usefulness of this input. The key point is that this was a subject expert with WP expertise so this panel member was a core part of all the conversations.

56. The panels (subject and main) do not reflect the diversity of the sector and it is important that the OfS create opportunities to bring in more academics and students of colour.

57. For much of the time together the panel explored perspectives that were shared across panel stakeholder groups and each stakeholder was not simply defined by their particular role. The employer panel members commented on a wide range of matters and their views, which were congruent with the rest of the panel, are reflected throughout.

Students' voice

58. The student panel members stressed the importance of clarity in relation to the purpose and audience for the submission. This links to the dual purpose of the submission as an assessment tool and as a public document. Some students felt that the submissions were dry and 'boring' and that they would do little to excite students about the subjects they might apply for. They were also concerned that the nuance of Bronze being above a quality threshold and resulting from benchmarked data would be lost on prospective students who could perceive Bronze as representing failure to meet sector standards.

59. There was a lot of discussion was about terms such as 'student engagement' and 'student voice'. For some the fact that five out of the six core metrics represented student-based views or student employment underlined that TEF is richly informed by student voice. Some of the students wanted to see something stronger in the submission that reassured them that students had an active role and agency within the subject. It may be useful to provide glossaries for these terms to clarify meaning in the context of future TEF iterations.

60. The panel recognised that its discussion had made few references to the role of professional and accrediting bodies and would strongly steer away from any obligation to include reference to this in submissions given the graduate employment and enterprise contexts for students in this subject.

Potential impacts

61. The panel was very concerned about the unintended consequences of subject-level TEF in relation to the creative arts education when there is a government focus on science, technology, engineering and mathematics (STEM), and a policy reluctance to talk about science, technology, engineering, arts and mathematics (STEAM). Factors that appear discrete are interconnected:

    a. 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 some further education and higher education providers experiencing dips in recruitment.

    b. 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.

    c. Creative arts and design has lower NSS averages against other subjects and whilst this is not a problem in TEF because the NSS is benchmarked it is a problem in any provider where the focus is on NSS league table positioning.

62. The panel expressed concern that 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.

TEF: Supporting the sector

63. If TEF is going to support the development of excellence it is essential that it is seen to 'give something back' to providers and subjects. This is particularly important given the increased workload and spread of burden across each provider associated with subject-level TEF. For this reason the Arts Panel agrees that a more nuanced profile-based approach to rating that was able to 'shine a light' on provider strengths via commendations would support the recognition and sharing of best practice. This approach would also address the panel's concern about the lack of innovation and experimentation in teaching excellence evidenced in the sample (with a few notable exceptions). A focus on imaginative innovation and enhancement would serve to surface excellence in a more dynamic way.

Final notes

64. The panel was uncomfortable with quotations from students, external examiners and employers (sometimes anonymous, sometimes with names) that may have been used in the TEF submission without permission. We view this as a repurposing of feedback that may have been given for one reason but is being used in another context to strengthen TEF submission.

65. We note the importance of TEF in relation to overseas recruitment and are concerned that the half-weighting of NSS leads to an effective double-weighting of graduate outcomes which do not include outcomes for international students.

66. The panel noted that postgraduate students applying for courses in subjects with a TEF rating are likely to assume that the TEF rating also applies to their course.

67. This report does not consider interdisciplinary issues because none arose in this sample.

68. The provider sample and the panel range of expertise and provider type was very strong but it is noted that the sample and the panel both lacked representation from Russell Group creative arts and design providers. It is recommended that this is addressed in the second year of the pilot.

Report author: Prof Susan Orr, Chair of the Arts Panel

Acknowledgements: I would like to thank Deputy Chair James Perkins and Quality Assurance Agency for Higher Education (QAA) TEF officer Derek Hamilton for their help in the production of this report.

Table 1: Arts Panel members

Chair	
Prof Susan Orr	Dean of Learning and Teaching Enhancement and Professor of Creative Practice Pedagogy, University of the Arts London
Deputy Chair	
James Perkins	Former Vice-President (Education), City University London Students' Union
Panel members	
Rob Brannen	Head of Visual and Performing Arts, De Montfort University
Prof Roni Brown	Deputy Vice-Chancellor (Academic), University for the Creative Arts
Stuart Cannell	Former Vice-President (Academic and Student Affairs), Belfast Campus, University of Ulster
Dr Dawn Edwards	Clerk to the Board of Governors and Head of Quality Assurance, Royal Northern College of Music
Cherie Federico	Director, Aesthetica Magazine Ltd
Prof Vicky Gunn	Head of Learning and Teaching, Glasgow School of Art
Jenny Hann	Assistant Director Higher Education: Curriculum and Quality, University Centre Weston
Chris Honer	Artistic Director, Library Theatre Company
Prof Emma Hunt	Deputy Vice-Chancellor, Arts University Bournemouth
Dr Mark Irwin	Dean of Higher Education, BIMM Limited
Eleanor Keiller	Guild President, University of Birmingham
Xenia Levantis	Students' Union President and Course Representative, Norwich University of the Arts
Joanna MacDonnell	Director of Education, University of Brighton
Prof Katie Normington	Senior Vice-Principal, Royal Holloway, University of London
Dr John Pymm	Dean of Faculty of Arts, University of Wolverhampton
Chloe Whittaker	Former Vice-President (Welfare and Education), Arts University Bournemouth

QAA TEF officer: Derek Hamilton

Business and Law Panel report

Executive summary

69. This report provides a review of the pilot process and outcomes for the Business and Law Panel. Most of the report represents the consensus of views from the whole panel. The section on student views has been written by the deputy chair and provides additional feedback solely from the student members of the panel.

70. Overall, we consider that the panel made robust assessments and that the ratings generated through holistic judgements accurately reflected the provision assessed on the basis of the process we were required to follow. We do have a number of recommendations we would ask to be considered in the development of the second stage of the pilot.

71. The panel considered that the subject categorisations were robust at the levels of business and management and of law. It was appropriate to include economics within the business and law panel where this reflected its institutional location.

72. The panel would like to see the introduction of a minimum cohort requirement in order to facilitate robust evidence-based judgements. We also recommend a review of the metric weightings to better reflect those aspects of quality (teaching, assessment and feedback, academic support and continuation) which are most directly impacted at subject-level. The panel concluded that the evidence base for student outcomes and learning gain was too narrowly focused on employment outcomes. We would like to see this broadened to allow greater consideration of the wider societal benefits attached to excellence in business and law higher education. In general, the supplementary metrics were not helpful to us in coming to our judgements and the nature of this evidence should be reviewed ahead of the next pilot. The contextual data was extremely useful.

73. Comparison of the two methods is necessarily limited at panel level but there was full agreement that the Model A subject-level submissions were superior to the Model B subject group submissions as an evidence base for our judgements. Many of the written submissions would have benefited from a more evaluative approach that triangulated evidence sources referred to. Stronger written statements tended to demonstrate how enhancement of learning and teaching was grounded in the institutional mission, addressed missing metrics and areas of weakness and included the student voice in ways that went beyond selective quotations from the National Student Survey.

Recommendations

74. The initial hypothesis at step 1a should be generated within the workbooks presented to panel members.

75. The half-weighting of the NSS scores should be reversed at subject-level and the double scoring of employment outcomes be reviewed.

76. Consideration should be given to the development of a metric which captures the social rather than the purely economic returns on higher education.

77. The training given to panel members should be reviewed in order to rebalance the coverage given to the quantitative and qualitative aspects of the holistic judgement.

78. Length and timings of meetings at post-pilot stage should be carefully managed in order that the same quality of decision-making is facilitated when working with significant volume.

79. Heuristics should be used at subject-level to ensure that panel time is used as efficiently as possible.

80. Value-laden terminology of Gold, Silver and Bronze may impact on the competitiveness of UK higher education in the international market.

81. The role of the widening participation and employment experts should be reviewed to ensure that maximum value is obtained from their input into subject-level deliberations.

82. The Model A subject-based approach for written statements should be adopted in both models and any subsequent variations thereon.

83. Providers should have access to support and guidance on producing subject-level statements.

84. The setting of a minimum cohort requirement for inclusion in subject-level TEF would help to ensure judgements are evidence based by reducing as far as possible the number of submissions with missing metrics.

85. Intercalated programmes should be excluded from subject-level TEF.

Generation of robust ratings

86. The panel was confident that it followed process and that this process led to the correct ratings as determined by the process at step 1a.

87. We note, however, that step 1a is purely a calculative process and that, to save time and possibility of human error, we would recommend that the initial hypothesis at step 1a is generated within the spreadsheets presented to panel members. This would also enable a greater focus on the judgemental aspect of ratings generation.

88. There was full agreement amongst academics and students that the half-weighting of the NSS scores and two scores for employment outcomes were problematic at subject-level. The relationship between staff and students at subject-level is mediated primarily through learning and teaching, and the quality of learning and teaching is captured in the NSS scores and continuation metrics. We would recommend that half-weighting of the NSS scores be reversed at subject-level and the double scoring of employment outcomes be reviewed. This would better reflect both the relative levels of agency which subject-level staff have over the different metrics and the lived experiences of students in their academic departments.

89. One possibility would be to replace one of the employability metrics with one that provides a measure of the return on investment in higher education in the form of social value and positive impact on the public good rather than solely in economic value and increase in personal capital. This could take the form of some measure of learning gain and could potentially build

upon the assurance of learning processes that are used in many business schools to measure programme learning outcomes[5].

90. The concept of 'best fit' was sometimes found to be problematic at subject-level with regard to the Gold descriptor's requirements and created a tension with the concern discussed below (paragraph 133) to be mindful of the consequences of the TEF ratings on the reputation of UK higher education in the international market.

91. Notwithstanding the issues raised above with regard to specific items within the dataset, the panel was generally confident that it was making effective use of the complete dataset, including the written statements, in order to arrive at holistic judgements. There was a concern, however, that the focus in the training on the metrics meant that the qualitative aspects of the judgement were underplayed. We would recommend that the training given to panel members be reviewed to give more emphasis to the review of the written statement and to include an overview of the state of the disciplines in the national context.

92. The makeup of the panel, particularly the diversity of its membership as regards provider type and discipline area, was considered to be a strength in enabling it to arrive at robust judgements.

93. The panel was content that the time allowed permitted it to come to unanimous or majority decisions on all cases it had to consider, with sufficient time for in-depth discussion of difficult issues. In particular, the two-day meeting schedule allowed for further review of cases where agreement was not initially reached by the subset of the panel who had read the case. The pace of the meetings also permitted others on the panel to review metrics and written statements in the course of deliberations where this was helpful. As a sector, we need to acknowledge that for us to reach sound and defensible judgements through a peer review process at subject-level will necessarily require a high level of academic and student input. We would recommend that careful consideration is given to how the sheer volume of business at post-pilot stage should be managed so as to allow the same quality of discussion and decision-making.

94. The flow of discussion was helped by the use of heuristics on the level of panel consensus to determine which cases would not require in-depth consideration. A sample of these heuristics was tested through comparison of their outputs with those following discussion and were found to match in all cases. We would recommend that heuristics be used at subject-level to ensure that panel time is used as efficiently as possible.

95. The panel was uneasy about the terminology of Gold, Silver and Bronze and how value judgements might be imputed from these with regard to individual institutions and to the UK higher education sector as a whole. Over 140,000 students from overseas, one third of all non-UK students, were studying business and law subjects in 2016-17[6]. As such, this subject panel was particularly mindful of the damage that could unwittingly be done to the reputation of UK

---

[5] See Association to Advance Collegiate Schools of Business (AACSB) Assurance of Learning Standards: An Interpretation https://naspaaaccreditation.files.wordpress.com/2014/04/aacsb.pdf.

[6] Higher Education Statistics Agency, 'Higher education student enrolments by subject of study and domicile 2016-17', https://www.hesa.ac.uk/data-and-analysis/students/what-study#.

provision relative to that of our international competitors. We would recommend more neutral nomenclature that underscores that all institutions participating in the TEF have met a rigorous qualification process.

96. The panel did not receive direct input from the widening participation experts as they were not present on the days when the panel was making judgements. Both the chair and the deputy chair observed how valuable their contributions were to the Main Panel discussions. We would recommend that the role of the widening participation experts be reviewed in order that optimal value from their expertise is obtained at subject-level.

97. The introduction of the 'No rating' category was extremely helpful as it prevented the panel being forced into awarding an initial hypothesis of Silver by default of lack of evidence where metrics were not reported.

98. The 'No rating' category was also applied in cases where business and management provision consisted solely of intercalated programmes, as the panel considered that only the NSS derived metrics were applicable, continuation data was not available and the causality of employment metrics was questionable.

Comparison of the models

99. The discussion in this report is limited to the part of the process that took place at the subject-level panel. Therefore, no comment is offered on the benefits and dis-benefits of the two models with regard to the relationships between subject-level ratings and institutional-level ratings.

100. The panel was unanimous in its preference for the single subject-level statements and considered that this approach worked best for all providers. The combined statements were generally confusing to read, with it often being unclear which particular subject was being discussed. The panel would strongly recommend the Model A subject-based approach for written statements. Submissions could be four rather than five pages.

101. The subject-based written statement could be usefully enhanced by a brief statement around institutional mission and ambition for the subject area.

102. The inclusion of 15-page statements for monotechnics at subject-level was not required or helpful and introduced inequities between providers. Single subject providers should be required to prepare the same four to five-page subject-based statement as other providers do for consideration at subject panel level.

Quality of the evidence

103. The quality of the written submissions in both models was variable, with those for Model B often being particularly difficult to follow at times because of a lack of clarity about which subject was being discussed at points within the narratives. There was also a lack of balance given to the different subjects within the statements.

104. Those written statements that were of high quality were more evaluative in nature (including telling us when things had not gone well) than weaker statements. The latter tended to be more descriptive and where impact was discussed it was often unclear about the extent of reach of

an intervention. Generally, the panel found selective quotations from external examiners' reports and NSS comments were not helpful unless they were triangulated with another source of evidence.

105. It is noted that the students' union boycott will have affected some institutions and this may be the reason for the student voice not being strong in some statements.

106. Submissions potentially envisaging the use of metrics from one subject by another needed to make clear the academic grounds for doing so. These might include the cognate nature of the programmes and the subjects they contain, for example, their use of shared modules and approaches to teaching, learning and assessment, shared teaching and physical resources, student forums.

107. The panel would recommend that greater guidance, such as the use of redacted exemplars, and training, particularly for institutions that have limited experience at subject-level of writing such documents, would be beneficial. The panel was mindful that there may be particular inequities between different providers of business and management, with larger and longer-established business schools possessing a local expertise because of their accreditation requirements. The majority of the panel was, however, not supportive of a template approach.

108. The panel was unanimous that an absence of data should not lead to a Silver rating at step 1a, since this was not compatible with the TEF descriptor for Silver. Nevertheless, where metrics are missing, putting the burden on the provider to prove they are not Bronze would be potentially unfair in the context of the concerns raised about value judgements made for different ratings.

109. Similarly, the panel was unanimous that reading across metrics from another subject was not appropriate and could create judgements that were misleading to users.

110. The panel would strongly recommend the setting of a minimum cohort requirement in order to reduce as far as possible the number of submissions with missing metrics. This is particularly important in the context of the business and law subjects where the rapid growth in the number of providers is likely to increase the volume of such cases over the next few years.

111. Whist the use of the dual hypothesis did help to ensure that some cohorts in minority mode were captured in the calculation and use of metrics, this was not so for the majority of providers with more than one mode of delivery. Even where metrics were reportable, sometimes splits were not because the numbers were too small. To address these gaps, it may be that widening participation and experience of minority mode students are focused on at the institutional level in the next TEF subject pilot.

112. The panel considered that supplementary data has potential but is currently of very little help. No panel member felt teaching intensity had helped them rate a case. Student satisfaction with the teaching intensity of their programmes is best captured through the existing channel of the NSS rather than creating a separate survey. The majority of panel members felt that the maps were of little value in rating cases, and misleading in at least one case. There was considerable scepticism about the reliability of LEO data and some panel members queried whether it was fair for some providers to have LEO data and others not.

Subject-specific considerations

113. Subject-level expertise was important in making judgements on activities and initiatives represented as innovative in written statements.

114. Subject-specific considerations came through more clearly in Model A written statements than they did in Model B.

115. There were no tensions identified between accreditations and profession, statutory and regulatory body (PSRB) requirements and the TEF processes being piloted.

116. Intercalated provision in business and management was problematic as continuation metrics were not generated and employment metrics cannot be related to the intercalated year. In most instances the effect of intercalated programmes will not be significant but where provision consists solely of intercalated programmes we would recommend that this is excluded from subject-level TEF.

117. Many business and management programmes emphasise entrepreneurship and students are encouraged into starting their own businesses. The employment metrics do not provide a good measure of outcomes for the self-employed. Therefore it is important that employer inputs are representative of both large and small businesses and the public and voluntary sectors as well as for-profit organisations. We would recommend that employer representation reflects the diversity of graduate destinations.

Segmented panel member views

Employers

118. The views of the employers were generally consistent with those of the broader panel and have been integrated into the report.

119. Employers expressed a concern that the three-part judgement of Bronze, Silver or Gold was not sufficiently variegated, a view that was shared by some other members of the panel.

120. There is a need to consider how we make most effective use of the employer time that is being devoted to this process. If it is to be the same as other panel members then we need to ensure that they are invited for times when cases are being considered.

Students

121. This part of the report has been written by the deputy chair based on a meeting of the student members of the panel.

What do the student panel members feel worked well in the panel this year?

122. As students on the panel, we felt that the panel treated all members with equality of expertise and equality of importance in panel discussions. There were no points throughout the subject panels at which student panel members felt unable to debate against their academic counterparts.

123. Furthermore, the chair made sure that throughout all discussions, the student panel members had ample opportunity to voice their perspective on cases, something that is

welcomed and should be replicated across all panels. This led to a sense of empowerment across all student panel members and the deputy chair of the panel.

What do the student panel members feel could be improved...

…for the next subject pilot?

124. The inclusion of an expert in widening participation on the business and law panel would be greatly appreciated by the student panel members. While we feel that in no part was the widening participation mission of institutions lost, this is a role that we feel should be consistent across all subject panels. We recognise that this was the initial aim of the TEF team but it was not possible to fulfil due to lack of suitable applicants.

125. In saying this, we agree the academic and student panel members each have their own widening participation passions and specialisms that were brought in full to panel members, and that widening participation forms a key part of the grade descriptors which were followed in the assessment process. But this specific additional role would aid with a holistic well-rounded view of widening participation from across providers all around the nations.

…for the subject-level TEF more generally?

126. Overall, we felt that the panel worked exceedingly well, and that there was a fantastic role that the chair played in ensuring that students and academics were treated as equals throughout the TEF process. We also felt that having an equal ratio of student to academic assessors on individual cases was a great benefit, and a credit to our TEF officer, who worked tirelessly to implement this. In order to build on this partnership of students and academics, we feel that an equal number of academic and student members on a panel would serve a great benefit to the conversation and parity of experiences that the panel consider.

On the theme of student involvement in the TEF process and the presence of student voice in the submissions:

How have the student panel members felt the process this year has gone?

127. There was a clear and noticeable lack of student involvement in the TEF submissions presented to the panel this year, and we feel that as a result, the student voice from providers was greatly weakened. We appreciate that this was partly due to the subject-level exercise being in its pilot year, and therefore the workload that would be placed on students or student representatives is unfeasible. But, this combined with the half-weighting of the NSS metrics, led to an overall shift towards an outcomes-based process, and we found ourselves at times questioning the amount the TEF was measuring teaching excellence.

How could the TEF specification and the Office for Students better support student involvement and the presence of student voice in the TEF process?

128. The student members of this panel have not reached a consensus on whether there should be a separate student submission to the TEF, or whether student involvement in the TEF submission as it stands should be a mandatory part of the specification. However, as a group we can agree that student involvement in the TEF submission adds a level of depth to how the learning and teaching strategies, so eloquently outlined and explained by institutions, work in

practice and delivery to undergraduate students. It was further noted that, while regularly the rebuttal for a separate student submission is the unintended disadvantage this places on further education colleges or institutions with a small or no student representative organisation, the further education student members felt this was in most cases an untrue conclusion.

Potential impacts

129. Subject-level TEF has the potential to encourage institutional managers to consider business and law performance against other providers, rather than focusing on comparisons with other subjects within the organisation, which does not always lead to helpful comparisons.

130. Given the size of business and law provision in most institutions, these subjects are less vulnerable to the threat of closure associated with a poor TEF outcome than might be some smaller subject areas. Nevertheless, depending on spread of subject-level outcomes and relative size of different subjects, an unintended consequence of subject-level TEF may be to shift institutional resource away from business and law provision and towards other subjects if improved TEF outcomes in these are considered more strategically important or more easily attainable.

131. Business schools draw large numbers of their students from overseas and Bronze ratings would be particularly injurious in a competitive, international context.

132. Many business schools and some law schools have significant postgraduate numbers. Given that TEF ratings are based solely on undergraduate provision, care will need to be taken on messaging of TEF ratings to prospective postgraduate applicants.

Other considerations

133. The panel gave consideration to what level of the CAH might be used in future iterations of subject-level TEF. There was a concern that splitting business and management into sub-groups would be to ignore the crossover between the constituent elements whereby the different elements are combined within programmes (for example, marketing being contained within management programmes) and their common cognate identity.

134. Additionally, analysis into smaller sub-groups would increase the numbers of cases with non-reportable metrics and split metrics.

135. The variety and level of expertise within the subject panel ensured that it was competent to deal with the totality of the discipline coverage contained within the returns for business and management.

136. We would recommend that the grouping for business and management remain at existing level and not be broken down into constituent elements at CAH3 (accounting, business studies, finance, marketing, management studies etc.).

137. Whilst the possibility of including economics in either the business and management or the social sciences subject groups may give rise to gamification, this risk needs to be balanced against recognising that the location of economics varies between institutions.

138. We would recommend that institutions be required to return economics within the subject that best reflects its positioning within the institution. For example, where the Economics programmes are taught out of the business school, it should be returned to the Business and Law Panel.

Report author: Prof Julia Clarke, Chair of the Business and Law Panel

Table 2: Business and Law Panel members

Chair	
Prof Julia Clarke	Pro Vice-Chancellor (Business and Law), Manchester Metropolitan University
Deputy Chair	
Josh Gulrajani	Former Students' Union Vice-President (Education), University of Essex
Panel members	
Matt Adie	Students' Union Vice-President (Education), University of Stirling
Andrew Bargery	Independent
Prof Warren Barr	Head of Department, Liverpool Law School, University of Liverpool
Prof David Boughey	Associate Dean (Education), University of Exeter
Deveral Capps	Dean of Leeds Law School, Leeds Beckett University
June Dennis	Principal, UK College of Business and Computing Ltd
Prof Tina Harrison	Assistant Principal, Academic Standards and Quality Assurance, University of Edinburgh
Dr Ashok Jashapara	Chair of Innovation Studies, Royal Holloway, University of London
Michael Olatokun	Students' Union Community Officer, University of Nottingham
Benjamin Phillips	Student President, Chichester College Group
Prof Elizabeth Smart	Head of Law, Sheffield Hallam University

Dr Sandra Summers	Academic Quality and Governance Committee and Student Senate Reference Group Student Member, the Open University
Ramita Tejpal	Dean of Higher Education, London South East Colleges
Juliette Wagner	Dean, Learning and Teaching, BPP University Limited

QAA TEF officer: Derrik Ferney

Engineering and Technology Panel report

General observations

139. A summary template at the start of the textual submission (particularly for Model B where different subjects were mixed together) would help the panel understand the courses, accreditation, student numbers and institutional structure. This should be factual and provide context. In addition, a more structured format or template for the textual submission would make the interpretation more consistent and accessible.

140. The panel struggled with the variable quality of the submissions without having access to the curriculum (see paragraph 169). In this panel the professional nature of the courses means that content underpins quality. As such many of the observations were on perceptions of quality rather than the true value and integrity of the courses.

141. The panel was divided over whether Silver is an appropriate 'default' position (as set out in the TEF specification). Is it better to give Silver as a 'neutral' rating in the absence of evidence with which to make a higher or lower judgement, or apply a more cautious rating of Bronze (or 'No rating') until there is sufficient evidence that quality exceeds the baseline?

142. The panel found a greater emphasis than expected on employment metrics, especially since these are a less direct measure of teaching quality than some of the other metrics being used. Using employment as the only measure of outcome is quite limited, and gives no indication of academic value added or learning gain.

143. The panel experienced difficulties considering the impact of one-year programmes (such as standalone foundation programmes) on the metrics, especially in relation to continuation and employment. It was felt that standalone foundation courses should be excluded from the TEF as they are not a 'destination' qualification.

Were robust ratings produced?

144. To normalise across panels a numeric score could be derived by a voting system at the panel. The normalisation and calibration would then be achieved by shifting grade boundaries to achieve an equitable distribution.

145. If rating cases against each other and moderating against a pre-defined distribution could bring greater equity across panels, the 'movable' boundaries might exacerbate mismatches with the TEF descriptors.

146. For the engineering panel the evidence around 'rigour and stretch' was limited in all cases. As the majority of the programmes require professional accreditation in order to practice or receive public liability insurance, the integrity of the curriculum has to be considered.

Comparison of the models

147. The panel experienced more instances of not reaching a consensus following discussion of Model A cases compared with Model B. Making a 'No rating' rating option available at Model A speeded up some discussions. In addition, the application of the 'No rating' moniker gave greater credence to those subjects that had sufficient evidence supporting a rating.

148. Metrics workbooks should be screened to exclude cases with limited data (in terms of non-reportable metrics and associated denominators) and given an automated no rating. Consideration could be given to a minimum amount of acceptable data. This would bring consistency across panels and reduce the burden for panel members.

149. Whilst metrics analysis could be further automated, panel members felt doing this part of the process themselves was necessary to generate 'the story' and understand the case.

150. Looking at other subject metrics in place of non-reportable metrics (principally under Model B) did not work for a variety of reasons, including differing initial hypotheses, a differing majority mode and lack of confidence in comparing to an 'imported' subject, which may well be located in a different faculty or school. The prospect of 'gaming' the submission was raised as a possible consequence of the Model B approach.

151. Could there be scope for giving commendations in particular areas or criteria?

152. As much comparability as possible with Research Excellence Framework (REF) principles – which have been developed and refined over time – would help build credibility for TEF within the academic community.

153. The burden for panel members was lower in Model A, partly due to accumulated experience but also by not having the subject 'detangling' requirement of the majority of Model B provider submissions, which in itself gave more confidence in the assessment.

154. Doing all the assessment work in a short space of time would help panel members get into a pattern – having the calibration event close to Model B assessments worked well.

155. In developing a possible hybrid model, the link between subject and the provider ratings is important. This would point to something based on Model B, with separate subject-level submissions. It would be important to establish the provider-level first before moving onto the subject-level assessments (perhaps using Year Two or Three outcomes in the first instance).

156. For the provider, producing five pages per subject every three years was not considered too burdensome. However, three years is quite a rapid cycle, and five or six years with a mid-cycle check on metrics for any significant changes could be better. A rolling programme could be another option. However, the rolling programme is likely to lead to an institutional industry developing around TEF, and potentially disadvantage small providers.

157. The panel felt it treated devolved nations consistently. The few instances where the provider type was evident in discussion was when aspects of institutional capacity and style were thought to have affected the quality of the submission.

Quality of the evidence

158. The overall balance of the metrics is very much towards student outcomes. Only two of the NSS-derived metrics relate specifically to teaching. However, students on the panel were explicit in describing the game-playing behaviour around submissions in the NSS, by voting more positively about institutional quality than more local subject-specific matters as 'they didn't want to damage the reputation of their institution.'

159. Of particular relevance is the selection of appropriate NSS metrics in relation to 'rigour and stretch'. Specifically, Section 1, 'Teaching on my course', and in particular Questions 3 and 4 are perhaps of most relevance – which the panel struggled to consider without evidence. The inclusion of 'The course is intellectually stimulating', and 'My course has challenged me to achieve my best work' would have helped the panel.

160. The reality of continuation varies considerably, and intentionally so, for different types of provider and provision. Is this really an appropriate metric and can it be correctly benchmarked? The continuation metric might be better based on completion or exit award only.

161. Employment aims and outcomes are also very different across different providers and provision. All are equally valid, but look indiscriminately better or worse in the metrics.

162. The inclusion of absolute values and splits for the core metrics was very useful. Supplementary metrics were not particularly useful. LEO data did not support the student outcomes and learning gain criteria as set out; it also reflected historic rather than current activity.

163. We are using proxy data, measuring what is measurable rather than what is directly relevant or which genuinely reflects teaching excellence. Teaching needs to be observed to be identified as excellent. Students need to be spoken to in person.

164. Ideas for alternative data and metrics are teaching qualifications (although there was a consensus that having a teaching qualification did not lead de facto to excellent teaching or education), external examiner views, a student submission.

165. The panel had some concerns over the very different student numbers in different CAH categories at the same level. Using the 35 subjects at CAH level 2 is an uncomfortable compromise, between seven subject panels and 155 separate disciplines.

166. The subject contextual data presented in the metrics workbooks was very useful. The data maps were in the main not useful. The derivation of the benchmarking data needs more explanation to the panel. The data was referred to frequently, but with some uncertainty over its provenance.

167. There needs to be a way to get providers to write better submissions which address gaps and issues in the metrics, most likely through a combination of template and guidance. Giving feedback on the submission would help improve the next one.

168. Panels need systematic information about professional accreditation, e.g. the percentage of programmes within each subject submission with accreditation.

Quality and robustness of the assessment process

169. The three-step assessment process is generally useful. Panel members felt that step 1b was the most influential, and provider submissions were not often persuasive enough to change the rating at step 3.

170. The panel had some concerns about how much attention, and so what relative weighting, were being given to provider submissions. More training on how to assess these submissions, and knowledge of how other panels were handling them, would have been helpful.

171. Institutional written submissions should be made available to subject panels to provide context – although this would give more work to subject panel members.

172. Some panel members would have liked greater understanding of how benchmarks are generated and what they mean, and so be able to better interpret metrics in relation to them.

173. Panel members would like more guidance on interpreting the TEF descriptors 'in the round', particularly concerning not 'ticking off' against every phrase or point, nor every element necessarily being at the same level, i.e. an holistic judgement without having to tick every descriptor. It was felt that the TEF descriptors required further work, in particular to make them representative of all types of provision.

174. Internal survey data was useful when there were gaps in the core and split data, but not when it contradicted the original dataset provided to the panel.

175. The use of selective individual student comments was largely ignored; those from accrediting bodies and external examiners had more credibility. Access to their reports would help calibrate the selection of comments.

176. The panel took a hard line on maintaining the position that the national quality standard threshold is Bronze. This level meets the national requirements, so institutions should be 'proud of this'. The observation that this would have implications internationally was the counterargument, where Silver is a more palatable norm.

177. Model A detailed discussions were faster and more focused, partly due to the more focused nature of the provider submissions but also reflecting the panel members' increased experience with the process.

178. The panel had some concerns over using metrics alone to identify cases for review (i.e. selecting exceptions) under Model A. It would certainly be better if the provider rating being used was the final one rather than the initial hypothesis.

179. Model B covered a wider range of disciplines in the round, but was too broad-brush to provide the specificity that would be of use to students. Aggregating information at a higher level would only compound the problem although it would lead to more robust statistical information (so long as the units of assessment remained within the same school or faculty – see paragraph 205).

Subject-specific considerations

180. CAH classifications – In the engineering and technology group, technology proved a problematic 'umbrella' and worked particularly poorly in terms of mapping onto provider organisational structures. Could technology be removed at level 2, and its level 3 subjects reassigned? It was noted that engineering at level 2 combines some very large cohorts with very different styles, e.g. civil engineering and electronic engineering.

Segmented panel member views

Academic

181. Rigour and stretch were not demonstrated in any depth in any submission (in contrast to professional accreditation submissions, where they are the pre-requisite). Unless this is addressed, the credibility of the process will be undermined.

182. The challenge of institutional buy-in to the credibility of subject-level assessment remains.

183. We need more exploration and analysis of the subject itself, whilst also balancing the burden of the overall exercise.

184. The Model B style of looking at everything potentially sends a better message in that all areas are under scrutiny, although it is too broad-brush to provide domain-specific information which is relevant to students.

185. Model A is likely to provide a localised impact and directly influence academic behaviours. Higher-level scrutiny makes the process less relevant to staff who are delivering the teaching.

186. We would prefer some kind of hybrid with Model A style five-page per subject submission.

187. TEF is an opportunity for teaching-intensive institutions to shine.

Employers

188. The subject-level treatment is important.

189. Written submissions need to include more specific issues that employers are interested in, such as transferable skills and depth of technical grounding.

190. Ultimately, employers want to know where it will be good for them to recruit.

PSRBs

191. The main risks of subject-level TEF concern credibility and any clashes of results with accreditations. The communication of outcomes will be very important. So the outcome should be associated with the level of the course (HND, BEng, MEng etc.). Some expansion of the overall rating descriptors to relate specifically to provision at different levels (i.e. Levels 4, 5 and 6) would be helpful. For example, ‘student engagement with developments from the forefront of research, scholarship or practice’ was difficult to translate into teaching methods across the range of provision under consideration, and some explicit description of the expected or likely differences would be useful.

192. A hybrid model of assessment would be preferable. Subject-focused submissions are essential.

193. The proportion of non-academics on panels could be higher, with more employers.

Students

194. We are not sure that prospective students would understand what Gold, Silver and Bronze ratings really mean at subject-level, and to what they actually relate, i.e. not faculty or programme. An observation from one student panel member, reflecting the findings of a

student union research project, was that some students would not apply to a 'Gold' institution because they were not good enough.

195. It would be helpful if the TEF process engaged with younger students – first-years and even A-level students – to understand their informational requirements.

196. How much of the panels' deliberations will be publicly available, any more than just a 'badge'. Could there be some way of communicating different types of Bronze or Silver with or without accreditation etc.?

197. We would like to see more student-friendly marketing of results, recognising that Key Information Set data embedded in Unistats is of little relevance to current students.

198. Could overseas students be included in the employability data?

Teaching intensity

199. Teaching intensity is not effective – the panel didn't use the data presented as it could not interpret it.

200. Contact time is not a measure of teaching excellence; it says nothing about student learning.

201. There was little information and guidance on what constituted effective 'online teaching' at each provider. Given the global trend in higher education to provide more material through virtual learning environments and massive open online courses, there was no consistent view on how they should be considered or valued. This needs to be resolved before the full launch of the scheme.

202. A measure of student engagement might be a more meaningful metric, such as NSS Question 4, 'My course has challenged me to achieve my best work', as might narrative rather than numerical information to support a learning opportunities criterion.

Potential impacts

203. The panel had concerns over whether prospective students will understand that the scope of the ratings is a broad subject area which covers a group of courses. Student interest is primarily at the course level, but the panel acknowledges that we need some level of aggregation. Student perceptions of what the results mean will need to be managed.

204. What messages are being given to wider audiences, e.g. employers, those overseas, and how do the TEF results compare with existing measures of UK higher education quality, e.g. professional accreditation?

205. There is real concern over how the outcomes will be fed into league table and subsequent funding sources (both home and international). The impact on international student mobility and governmental accords may be affected.

Report authors: Prof Nick Lieven, Chair of the Engineering and Technology Panel, and Helen Uglow, QAA TEF officer

Table 3: Engineering and Technology Panel members

Chair	
Prof Nick Lieven	Professor of Aircraft Dynamics, University of Bristol
Deputy Chair	
Lewis Cleminson	Students' Union Education Officer, Southampton Solent University
Panel members	
Prof Hassan Abdalla	Pro Vice-Chancellor (Learning and Teaching) and Dean of College, University of East London
Olumayowa Ayodeji	Engineering Business Management Course Representative, Coventry University
Prof Gillian Cooke	Head of Teaching, School of Engineering, University of Warwick
Dr Philip Hanna	Director of Education, Queen's University of Belfast
Dr Alan Hayes	Director of Teaching Computer Science, Associate Dean (Learning and Teaching), Faculty of Science, University of Bath
Prof Alan Kwan	Professor, Deputy Head of School (Teaching), Cardiff University
EUR ING Graham Orchard	Director, Engineering Synthesis Consulting Ltd
Prof Shushma Patel	Director of Education and Student Experience, London South Bank University
Deborah Seddon	Policy and Standards Adviser, Engineering Council
Uyanahewage Viran Silva	Aerospace Engineering Course Representative, University of Sheffield
Prof Rachel Thomson	Pro Vice-Chancellor (Teaching), Loughborough University
Claire Wallace	Senate Reference and Student Experience Committee Representative, the Open University

Prof Sean Wellington	Associate Dean Strategy and Development, Oxford Brookes University
Kate Williams	Head of Higher Education, Greater Brighton Metropolitan College
Sarah Woolham-Jaffier	School Officer and Class Representative, Heriot-Watt University

QAA TEF officer: Helen Uglow

Humanities Panel report

Executive summary

206. This report outlines the experiences of the Humanities Panel in the TEF subject pilot 2017-18, and makes a series of observations and recommendations. The panel worked well and collaboratively, and student, academic and employer representatives all confirmed they felt engaged and included. The panel did note that greater diversity in membership, especially with regard to ethnicity and broader experience of working across an increasingly differentiated sector, would be important for any substantive subject-based exercise.

207. The Humanities Panel valued the balance between metrics and submissions, and was committed to reaching a holistic judgement on ratings. However, the quality of submissions across the two models was highly variable, potentially compromising the panel's capacity to come to genuinely holistic decisions. We advocate clearer and more specific guidance to providers in future to ensure that submissions are effective and have the potential to move ratings up or down from what can be a 'sticky' initial metrics-based hypothesis. A particular focus for guidance should be the effective involvement of student voice in submissions at all levels.

208. The panel notes a series of challenges and limitations for both the current Model A and Model B, and does not believe that either, as currently configured, could provide a feasible, robust and scalable option for sector-wide subject-level TEF. These limitations are not insurmountable, but genuinely rigorous and well-evidenced ratings will not be achieved without some re-engineering, and without additional cost and burden to providers and assessors. The panel favoured the subject focus of Model A submissions over the subject group submissions from Model B, with subjects feeling closer to the intellectual identity of staff and students. However, assessing only exceptions in Model A meant some subjects would not be considered at all, reducing their motivation for enhancement. Any future model should exclude exceptionality and subject group submissions.

209. Regardless of model, the panel found the teaching intensity data unusable, and strongly supports its removal from TEF. The panel argues for the introduction of additional contextual data on the type of qualification students arrive with, and on the programmes in each subject at each provider, with student numbers for each. There was concern about the reduction in emphasis on the experience of current students through the metrics, in favour of essentially historical employability data. A case can be made at subject-level for having a different metrics profile to that at institutional level, with a reduced focus on employability. In humanities, many students study more than one subject, and it will be important for this experience to be captured in future rounds of TEF; this may involve some changes to the subjects recognised, and certainly requires more guidance for providers on commenting explicitly on such cohorts. The panel strongly supports the inclusion of a narrative element, as well as a rating, in the information returned to providers. This would make outcomes useful to providers, and allow commendations and recommendations to be expressed.

Panel working

210. The Humanities Panel worked effectively and was extremely ably supported by an exemplary TEF officer. The general consensus was that panel members contributed on an

equal basis, with every member bringing their own insights and experiences. Members did not feel constrained to champion a specific perspective, though expertise of all sorts was respected, and all voices were heard. It is therefore unnecessary to provide a separate section on segmented panel perspectives. The panel chair and deputy chair found it particularly reassuring that student and employer members confirmed they felt they had been included and attended to on a completely equal basis. Student members did feel particularly strongly that the focus of TEF on outcomes and particularly employability might have gone too far, and were concerned about the half-weighting of NSS metrics and the consequent downgrading in importance of the experience of current students. However, while this point was put forward by student panel members in particular, it was supported by the panel as a whole – while recognising that the NSS could be volatile because of boycotts, industrial action or simply low participation.

211. The panel benefited considerably from the presence and advice of an employer representative, and felt a perspective from outside the sector was important, refreshing, and brought focus back to the external reception and perception of TEF outcomes. The panel would also have appreciated a specific WP member. On the basis of the pilot submissions, there will be less call for a PSRB perspective, though the panel did encounter a small number of programmes (e.g. in communications and media) with accreditation. The panel reflected on its use of subject specialists, which is obviously more of an issue for Model A, and did reopen one case where on reflection we felt we had not sufficiently prioritised the view of a subject specialist. The panel made rigorous attempts to be consistent in factoring in subject specialist views: how this works in future will naturally depend on the model or models tested in the TEF subject pilot 2018-19.

212. While the panel had a good balance of men and women, and representatives from all the devolved administrations, we acknowledge that our membership was not sufficiently diverse in terms of ethnicity, or optimal in terms of current expertise from across the sector. This was a particular issue in terms of experience in further education colleges and alternative providers, but the panel was conscious of increasing differentiation across the sector and of the distinctiveness of provider missions more generally. Numbers of further education colleges and alternative providers submitting to the humanities panel were very small, but if subject-level TEF is to be meaningfully scalable in future, it will be vital to find ways of encouraging greater diversity in applicants for panel membership, both in terms of sector experience and protected characteristics.

Were robust ratings produced?

213. The panel valued the balance between metrics and submissions in reaching an overall, holistic assessment. While some panel members started out feeling somewhat insecure about using the metrics, by the end there was consensus that the three-stage process remains effective, and support for retaining a non-automatically generated stage 1a to encourage assessors to engage fully with the metrics and reflect on contextual information about the provider.

214. However, there was a sense that using the submissions could be a challenge, and that there might have been some over-reliance on the metrics. Once an initial hypothesis (or subject-based initial hypothesis at provider-level) was reached, this could be quite ‘sticky’ and difficult to depart from. This may in part reflect the emphasis in training on using and

interpreting the metrics. While this was seen as inevitable in a year with new metrics being introduced and a significant number of new assessors, the panel would appreciate greater emphasis in future training on using the submissions to reach a holistic judgement, and on widening participation issues. The panel also recognised a tendency to use submissions more readily to 'promote' a case to a higher rating, and greater reluctance to allow even a very poor submission to award a rating lower than the initial hypothesis. There was nonetheless general agreement that the submission had to be equal in principle with the metrics, and should be able to play an instrumental role in determining the final rating for a provider. This year, submissions did vary very considerably in quality. Although this perhaps reflected the tight timescales afforded the pilot participants, the panel would support clearer guidance to providers on writing submissions (see paragraph 242).

215. There was a concern that we arrived at too many Silvers in both models. While some reassurance was afforded by the fact that the spread of end results was similar to that for TEF Year Two, the panel was uneasy about the number of factors leading to an initial 'default Silver' at metrics stages. This may tend to an interpretation of Silver as 'average' or threshold level, whereas this would instead fit the descriptor for Bronze. In particular, there was substantial concern, given the greater tendency for unreportable data at subject or even subject group level, and especially in smaller providers, that a subject with little or no metric detail and very little additional explanation in the submission could end up defaulting to Silver. The panel felt that in some such cases, the more appropriate approach was to move to Bronze after consideration of the metrics, with the onus on the provider to argue its case in the submission to raise this. We concluded that, unless a creditable way could be found of handling very small subjects with little reportable data, this might be a serious issue for the coherence and credibility of subject-level TEF. Naturally, such cases were also extremely time-consuming for the panel. Subsequently it was confirmed that such cases could be given 'No rating'. This was warmly welcomed by the panel as resolving the immediate issue, but recognised as not providing a satisfactory solution in the long term, partly due to concerns over the possible public perception of 'No rating'.

216. While further detail is given at paragraphs 228-237, the panel was in agreement on the main issues for each model. For Model A, the major obstacles were the assessment only of exceptions, and the number of small subjects with no viable data. For Model B, the burden of assessment was substantially greater, and subject group submissions were often of lower quality, probably reflecting the fact that subject groups are not 'real' units for students or academics in many cases.

217. The panel considered whether different metrics, or a focus on different aspects of quality, might be appropriate at institutional and subject-level, and would encourage further debate on this matter. TEF at subject-level might reasonably focus on teaching excellence, with less emphasis on student outcomes, which may be better considered at institutional level. This would also inform and indeed require a different focus in submissions at provider and subject-level; some initiatives (e.g. mental health support or careers advice) may properly take place and be reported at provider rather than subject-level, for example.

218. The panel appreciated the contributions of the main panel, and the overlap in membership as panel chairs and deputy chairs are also main panel members. There was a recognition that the main panel could and did take a cross-subject perspective and provide advice on difficult issues such as unreportable data. The panel was keen to hear more about the analysis of

results across panels, and evaluation of the reasons behind potential variation between subject panels. While recognising that the main and subject panels needed to stay distinct, there was also a strong sense that there should be a greater link to the provider-level submission and assessment when making subject-level ratings.

219. Overall, while the panel was confident that a rigorous approach had been taken to reaching ratings, we were uncomfortable that the rating alone would be all the information received by pilot providers. The panel accepted that a major motivation in this subject pilot was to explore options for a feasible, scalable model, and that pilot providers understood the nature of feedback this year. Nonetheless, we felt future feedback to providers should crucially be usable as well as well evidenced. A statement of findings, as in TEF Year Two, would present one option for signalling areas of commendation and recommendation to providers, and provide reassurance for some of those rated Bronze, for instance, that they were absolutely on the right track but with insufficient time for their initiatives to have shifted metrics. Some text-based feedback would also supplement ratings in allowing for 'high Silvers' and 'low Silvers' to be identified; without an option of this sort, some panel members felt that the current three ratings should be shifted to a five-point scale.

Quality of the evidence

220. There is a clear issue for both models on handling subjects (or indeed even providers) with low numbers or substantial unreportable data. The panel considered whether there should be a cut-off point for seeking a subject rating: with fewer students, a provider would need to aggregate up to subject group or provider-level. As noted at paragraph 217, returning a 'No rating' resolved the immediate issue for the panel, but does not address the underlying issue of small populations.

221. In terms of the new secondary metrics, the panel agreed that the 'grade inflation' data encapsulated a reasonable question to ask of providers; some did provide insightful outlines of the factors behind changes in their patterns of attainment. We would strongly prefer another title for this dataset, however, since 'grade inflation' carries a clear negative implication, and there can be very good reasons for an uplift in outcomes: if a provider has successfully closed an attainment gap between students from different groups, we would surely not expect an uplift in one group to require a reduction in good honours from the previous higher achievers.

222. On the other hand, the panel had no faith in the teaching intensity data and made a unanimous decision early on in the process that it was simply unusable. The general comparison document produced by the TEF team tells us that there are variations by discipline but this is well known; and referring to these subject-specific patterns only normalises what is done now, without allowing any questions as to whether there should be such variation across disciplines in the first place. In any case, student dissatisfaction with contact hours can reasonably be expected to come through in other NSS responses. The panel felt it was exceptionally simplistic to take a view that it is always better to have more contact time, and considered that it matters much more what is being done in that time than who is in the room (so, a professor reading a textbook to a class would not be better value than an hour with an early-career colleague who is an expert teaching practitioner). We also recognised an imperative towards higher education developing independent learners, especially at later stages of study; and were concerned that yet another survey of students would threaten the

volume of responses to NSS. Most importantly, we simply saw no place for teaching intensity, as an input measure, in an outcome-focused exercise.

223. The panel felt that the contextual data relating to student tariff was partially helpful, but that it would be at least as helpful to know the qualifications students come with (e.g. BTEC, A-level, or Access) as this could impact on thinking around widening participation and retention. For example, if many students studying an essay-heavy subject come from a BTEC background, then the institution may have more support work to do on transitions. The panel would also have valued supplementary data on which programmes of study are offered and how many students are on each. In Model B in particular, there was sometimes no mention in a provider submission of a particular subject, and some subject groups are very broad. For example, a communications and media return might include journalism, or games design, or both, and the experience of these students (and the modules they might choose alongside their major subject) could be very different.

224. There were still some queries and concerns about how effective the national benchmark for employability is and some panel members remained unconvinced of how to interpret this. The panel did use the LEO data but questioned whether it was too old to be of direct relevance to providers' current strategies, and whether the revisions to DLHE might mean there was less of a need for both these datasets in future. Some submissions made little mention of minority-mode part-time students, although there might be substantial numbers. Where part-time employability patterns were very different from the full-time cohort, it could be challenging to interpret why this should be the case: was the provider very good at encouraging part-time students into employment, or had they started their programmes in a job and kept it throughout? Motivations for study can also be very different across different populations, and this is especially salient for certain providers – 'leisure' higher education or evening classes have an important place but can sit uneasily with employment data. The panel would have favoured an additional age group category, say 50+, to highlight students who might be studying with less of a focus on employability.

225. Finally, panel members noted that it was difficult to interpret the maps, in particular the national scatter of jobs by sector. Institution-specific maps could be helpful, but were inconsistently referenced in the submissions. A suggestion was made as to whether the maps could be linked to the LEO data and whether this might improve the effectiveness of both resources. However, if they are to be retained, they need to be made accessible for colour-blind panel members.

Comparison of models

Model A

226. On the whole, the panel felt more comfortable with Model A, where submissions were much easier to navigate, and it appeared that institutions knew better how to present these. We felt we were hearing more authentic voices of discipline areas in Model A; and the load was clearly lighter and more manageable. The central flaw, however, was that we did not hear any voice at all from those subjects which had not been identified as exceptions. If TEF is to be an effective motivator of enhancement, there is little or no such motivation for subjects which are not exceptions and for which no submission therefore needs to be made. Nor could non-exceptions evidence greater excellence than the metrics would indicate.

227. Model A also missed important parts of the narrative, for example how joint programmes are delivered (noting that not all Model B submissions did this either, but there was the opportunity to). In Model A, decisions will be made about a large number of subjects without the subject specialist panels even seeing them or knowing they exist. The panel felt this was not an appropriate design for a public-facing model.

228. The panel therefore saw Model A as essentially incomplete – we did not have a full set of comparisons. We did not see all the subjects from single providers, or all the same subjects across providers, and we did not know the provider rating to which the subjects we did see were exceptions. Providers might have found it easier to write subject submissions if they could have been transparent about what they were exceptions to; and the panel would have benefitted from sight of the full provider submission for better awareness of the context for the subject-level students, as some initiatives are properly focused at institution level (such as perhaps mental health support). As matters stood, some providers chose to write a primarily institution-level submission with a section on the subject, while others focused on subject-level and seemed to assume that we had read the provider submission, which was not the case.

229. These issues made Model A too compartmentalised, making comparisons challenging and losing a holistic view. The drivers for this approach may be objectivity and independence, with one panel looking at the provider-level and another at subject-level; and of course this is important. However, the panel considered this analogous with the difference between full double-blind marking of student assessments, and moderation on the basis of knowing what the first marker has given and why. Most universities have moved to a moderation approach without compromising standards, and on the same basis a provider submission and rating could have been made available to subject panels for Model A without compromising objectivity.

230. The role of subject specialists on the panel was much more pronounced in Model A than Model B, and this was helpful and insightful – but if all subjects per provider were included, and we required full subject coverage on the panel, this would clearly add substantially to the cost of TEF. At the same time, the panel felt some discomfort that Model A did lead to a definite focus inwards onto the single discipline, whereas we know that many humanities students have a joint honours experience or routinely take modules across disciplines.

Model B

231. Model B did potentially provide the panel with the capacity to reach a more convincingly holistic view, given additional contextual knowledge. Subject group submissions could be beneficial in aggregating information, but could also lead to subjects being ‘hidden’ and not mentioned. While it is not possible to second-guess provider motivation in excluding a subject from a subject group submission, the panel felt that any future iteration or variant of Model B must include a direction to include at least some mention of every subject. This would resolve the problem encountered in Model A, where focus is on single subjects and the frequent experience of joint honours students can be excluded. It would also assist in cases where there are gaps in the metric data which are not explained through the submissions. In the same way, the comparability of subjects across submissions in Model B was often helpful, both for the same subjects at different institutions, and different subjects at the same institution. Model A often did not allow this and made judgements, comparisons and borderline decisions more challenging.

232. In Model B the panel tried, and to an extent liked, working in three smaller groups, each group taking one provider with about four subjects, before coming to a final conclusion among a larger panel. This was not workable in Model A for reasons of time, but for Model B did replicate some of the advantages those involved in TEF Year Two had experienced from working in smaller and then larger groups, and allowing decisions to be interrogated and revisited.

233. The panel was concerned by the possibility of gaming when providers shift an area into or out of humanities, or into or out of a subject group in Model B. We understand that the provider needs to provide a rationale for why subjects are being moved, and would have appreciated seeing this rationale at subject panel level.

234. Some aspects of Model B are also very granular, with small populations meaning that any splits will automatically be unreportable. This has serious consequences, as it is impossible to tell whether an institution is really supporting its populations with protected characteristics, or minority-mode, typically part-time students. TEF currently looks like a process which will allow us to evaluate what is being offered to widening participation students and students with particular protected characteristics. However, bringing it down to subject-level, with so many small programmes and subjects (at least in the humanities) and hence so many non-reportable metrics, may actually mean we lose the capacity to fully evaluate widening participation considerations.

235. While Model B was time-consuming, and there was some inconsistency with a lack of information on certain cohorts, it still felt more comprehensive to present to the subject community than the current Model A. The panel felt that it was too early to fully evaluate Model B, since some of the issues highlighted could be quite easily resolved. In particular, improvements in provider submissions would contribute to reaching more robust ratings. On the other hand, there is a more fundamental issue with the use of subject groups, which do not relate so closely to staff or student identity and learning community experience as subjects. The panel was concerned that requiring submissions on subject groups involved reifying units which are more about institutional structures and data returns than patterns of learning and teaching. While the voices of subjects came through in Model A in an authentic way, the subject group submissions were patchier and did not always demonstrate ownership of the data or real knowledge of the students. This is the main drawback of Model B.

Subject-specific considerations

236. The humanities panel encountered a number of issues around the Common Aggregation Hierarchy. Many, indeed probably most students in humanities either have a joint honours experience, or study combined subjects because they are able to choose modules from outside their major subject. The panel was very concerned that we did not have enough detail about students in joint or combined honours provision. This has obvious consequences for how providers return ‘combined’ students to the Higher Education Statistics Agency, which would need to be attended to in future rounds of TEF. In particular, the panel encountered a number of problems with the two combined and general studies subjects, and the humanities and liberal arts (non-specific) subject. There are numerous cases where it is difficult to identify a specific cohort of students under these headers, who are following anything like a common programme and might be thought to share anything like a common experience. While there were some good examples describing this provision, in other cases there is little or no

description at all, while in yet others there appear to be two or more distinct cohorts with very different characteristics and experiences. We might see a case for retaining one of these categories at Level 2, for instance to allow for those providers who really do have, e.g., a liberal arts programme, an area where there has been a significant increase in provision over the past several years, or for general degrees in Scotland. However, we would encourage a requirement for providers who choose to return students under this header to provide some rationale for their inclusion as a separate and identifiable cohort.

237. Beyond the 'basket' general categories, the panel also queried whether 'information services' would be better in computing than in communications and media and humanities. There is already an 'information studies' element under computing which may overlap. Likewise, there are some questions around the languages, linguistics and classics subject group. Neither classics nor linguistics is big enough to be a subject group on its own; but the panel noted that in many cases, linguistics is part of or in the same unit as English rather than languages, or falls within social sciences. In the case of classics, there may be a case for including it with history and archaeology (there are also a number of providers with separate programmes or units in ancient history, and classics and ancient history might be a viable subject at level 3). We recognise that removing linguistics and classics might leave a very small subject group with only languages, and this would need further consideration. Finally, there was some discussion around film studies, which is developing as a separate discipline; and about the fault line between humanities and creative arts.

238. The core issue here is that some of these subject groups will map very differently onto institutional structures, which could be misleading when results are published. Alternatively, in some cases students in one subject area are spread across departments or schools, which may receive different ratings. This may be another argument against the use of subject groups; or in favour of greater flexibility for institutions to choose the subject groups that fit their own structures best, though this in turn might compromise comparability.

239. Beyond the Common Aggregation Hierarchy, the panel was concerned about subjects which may be new, even though the provider is well-established; and about subjects which are being taught out. In the former case, we felt a Provisional rating for the first three years would be more appropriate than seeking to reach a rating, or giving a No rating.

Feedback on submissions

240. The panel was very clear that there needed to be improved and more specific guidance on submissions at all levels. This is particularly true if we are to maintain the importance of the holistic final judgement, which does entail that the quality of the submission could be instrumental in moving a provider or subject up or down between ratings.

241. Some of the submissions the panel saw in both models were excellent. These were generally characterised by engaging fully with the metrics and data, and obviously knowing the students and understanding their experience. However, there were also poor submissions which missed opportunities, failed to provide evidence of impact, or were simply extremely thin and poorly written. While the panel accepted that time was short this year for providers, we cannot simply assume that everything will come right if there is a bit longer in the cycle to write the submission.

242. The panel did not necessarily favour a template solution, and was keen not to stifle creativity or to filter out the good and positive elements of submissions that capture the essence of a place. On the other hand, we recognised that some providers may have less resource and less experience, and did not want to risk disadvantaging smaller providers or those which had not engaged with an exercise of this sort before.

243. One aspect of any potential guidance on submissions involves the optimal inclusion of student voice. The panel debated whether there should be a separate student submission, but equally noted that some of the most effective submissions had the student voice fully integrated. The crucial issue is to ensure that submissions make it clear how students have been involved. The panel's student members would be happy to be involved in producing some guidance for institutions on what would count as Gold, Silver or Bronze student engagement.

Potential impacts

244. While the panel clearly stood behind the ratings it had reached, there was significant discussion of unintended consequences.

245. First, the panel recognised that it was dealing with a number of small and potentially vulnerable subjects, sometimes in a group of larger and potentially better resourced ones, and was therefore seriously concerned that TEF might end up playing a role in reducing student choice. Panel members recognised that a small subject rated Bronze, in a subject group which is mainly Silver or Gold, may find itself under threat. If a narrative to providers were available, the panel could make the case that with support and investment, which the provider may transparently be making in adjacent subject areas which are doing much better, this area would have an excellent chance of thriving. Instead, we were acutely aware that we might be disincentivising existing and emerging small and specialist provision. Any potential aggregation of this effect up to discipline level across providers is a particular concern for humanities given that there are some subjects, like languages, which are at risk of closure at a national level. In this connection, the panel also noted that there is a highly significant gender bias in some humanities subjects. If some of those were at risk, we would particularly be reducing choice for women.

246. Similarly, the panel did not wish to stifle innovation in providers. We noted that continuation figures are sometimes poorer where many widening participation students, or students trying out higher education when they have no family experience, come into an institution; or where a formal qualification is not the primary aim of a cohort of students. If this leads to a lower rating, we were concerned that this might discourage providers from introducing or continuing initiatives which help local communities.

247. There was also some concern that large providers with excellent intake tariff and strong students are at risk of being penalised because very high absolute value markers do not fully compensate for a perception that they 'should have' a flag for a metric, even though this principally reflects an extremely high benchmark. While the panel did double-check such cases, we had a sense that we might be better at compensating providers with challenging cohorts and difficult regional factors in terms of employability, even where they may not have provided good evidence of mitigation through their submissions.

248. Finally, the panel was concerned that the move towards a more outcomes-focused TEF might mean perverse incentives could come into play. If TEF becomes known as a student outcomes exercise based on five-year-old data, it will become harder to convince colleagues that we should be focusing on improving the student experience and teaching quality for current students. This in turn would jeopardise one element of TEF which the panel felt was important and well-evidenced, namely the encouragement of an institutional drive to improve engagement with education in providers of all kinds.

Report author: Prof April McMahon, Chair of the Humanities Panel

Table 4: Humanities Panel members

Chair	
Prof April McMahon	Deputy Vice-Chancellor (Education), University of Kent
Deputy Chair	
Peter Cowan	Vice-President, the Open University Students' Association
Panel members	
Prof Sean Allan	Professor of German, University of St Andrews
Harry Anderson	Former Guild President, University of Liverpool
Dr Mark Bradley	Associate Pro Vice-Chancellor (Education and the Student Experience), University of Nottingham
Prof Phil Cardew	Deputy Vice-Chancellor (Academic), Leeds Beckett University
Dr Elaine Fulton	Reader in History Teaching (higher education), University of Birmingham
Mathew Gillings	Former Students' Union College Officer, University of Lancaster
Toby Gladwin	Former Guild President, University of Exeter
Liz Harris	UK Lead on Sourcing and Recruitment Marketing, International Committee of the Red Cross
Prof Jan Jedrzejewski	Professor of English and Comparative Literature, University of Ulster
Mary Ann Kernan	Associate Dean (Student Experience), School of Arts and Social Sciences, City, University of London

Prof James Knowles	Vice-Principal and Executive Dean of Arts and Social Sciences, Royal Holloway, University of London
Dr Melanie Prideaux	Associate Professor of Religious Studies, Deputy Pro-Dean for Student Education Faculty of Arts, Humanities and Cultures, University of Leeds
Ruth Stoker	Director of Teaching and Learning for the School of Music, Humanities and Media, University of Huddersfield
Hannah Todd	Vice-President of Education, University of Glasgow
Barnaby Willis	Former Students' Union Vice-President, Cardiff University
Prof Tim Woods	Director of the Institute of Arts and Humanities, Aberystwyth University

QAA TEF officer: Rafe Smallman

Acknowledgements: I would like to thank Deputy Chair Peter Cowan and Quality Assurance Agency for Higher Education (QAA) TEF officer Rafe Smallman for their help in the production of this report.

Medical and Health Sciences Panel report

General findings and key messages relating to the functioning of the models

249. Both models were implemented successfully in accordance with the specification identified for implementation. Assessors felt confident that final decisions were robust, despite time restrictions in the assessment process. In Model A, strengths related to the subject-focused written submission and in Model B, strengths related to the review of all subjects. The panel valued written submissions and suggested that with increased guidance for both providers and assessors, these could strengthen in quality and contribute more effectively to the holistic judgements made. Associated with this, review of the TEF criteria could enable greater clarity of outcomes for students and a more useful model for holistic assessment. The panel identified that for some providers, neither model was fully suitable. This was either because the specialist mission of the provider did not accord with the metrics, or because the subject included a large number of small specialist study courses (e.g. subjects allied to medicine not otherwise specified), or because the nature of the course meant insufficient data was available to make a confident judgment. The panel recommend these cases are examined further and provision made to address their needs within future assessments.

Subject specific findings and key messages

250. The panel identified that TEF should be reviewed to ensure alignment in respect of professional 'fitness for work' within regulated professions and assure credibility of the exercise. This is particularly important where measures may be influenced by regulatory requirements (e.g. teaching intensity) and where measures are not required for graduation (normally accredited as part of course requirements) but are additional for professional registration (e.g. licence examinations), but fall within the immediate post-graduation period. Finally, in some subjects reviewed, the panel recognised that employment may not be of key concern for potential students due to high workforce demand. In these cases, the panel suggest that the current metric weighting of employment outcomes as measured by two fully weighted Destination of Leavers in Higher Education metrics (and supplementary Longitudinal Education Outcomes) data appears high. Greater focus upon student experience as identified by the three National Student Survey metrics (currently half-weighted) and full exploration of other potential measures of employment and teaching quality could offer more emphasis on teaching excellence and a greater benefit for students.

Conclusions

251. This report concludes key points regarding opportunities for developing a model for teaching excellence at the subject-level in medicine and health sciences subject areas. It makes general observations which may be collated alongside the observations of other subjects. Findings are offered in the spirit of constructive and collegial support for the enhancement of teaching excellence by the panel.

Introduction

252. The development of a new tool for the enhancement of learning and teaching at the subject-level within medicine and health sciences has represented a unique opportunity to contribute to the shape of future developments in the subject area and the panel would like to

thank the OfS for this opportunity. The subject group as drawn from the HECoS CAH at level 2[7] includes seven core subjects within the grouping, but with many smaller subjects included. This is an important context, as many (but not all) subjects included within the pilot in this subject group required some accreditation or legislated regulation from PSRBs in addition to institutional requirements.

253. The full range of subjects included for consideration should be viewed within the HECoS CAH for completeness. However, to offer a flavour for this report, examples included classic university subjects such as medicine and dentistry as well those more usually associated with delivery within a further education college (e.g. foundation degree programmes for allied health professionals and in subjects allied to medicine, complementary therapies and counselling). Programmes such as nursing and allied health (e.g. physiotherapy and sport sciences; pharmacy, pharmacology and toxicology) were also included. Many subjects were professional, vocational or both, including significant practical training as well as theoretical studies, and most required a significant element of 'caring' for others.

254. The subject panel, recruited to evaluate the application of two models (Model A and Model B) proposed within the Department for Education's TEF subject-level pilot specification (2017)[8], reflected the broad scope of the subjects. Assessors were drawn from the student body including new graduates, academics and colleagues from Scotland and Wales. There was representation from employers and people with PSRB experience. Panel meetings were attended by colleagues with widening participation experience who observed a number of panels, and TEF team and main panel members observed on a number of occasions. The chair and deputy chair collated key points on each occasion that the panel met for feedback to the TEF team. In this way, communication around the process and activities of the panel was maintained in a continuing dialogue. Panel review of submissions was conducted in accordance with specific requirements identified by the TEF administration team. This assured key aspects such as confidentiality were addressed and key rules for undertaking the process could be clarified.

255. This paper reports on critically constructive observation and formal feedback discussion undertaken by the Medicine and Health Sciences Panel during the assessment and two pilot meetings held. It addresses the process and evaluation of using two models, using either a top-down (Model A) or a bottom-up (Model B) approach. For coherence, the report reflects the activities of the panel in the sequence in which they were undertaken and so where the models are divided, Model B is addressed first in each section.

Robustness of ratings

256. Model B ratings were reported by the assessors to be accurate, with a high level of confidence in the final decision achieved (albeit, on occasions, with considerable discussion). Exceptions included the following.

[7] HECoS (2018), https://www.hesa.ac.uk/innovation/hecos.

[8] https://www.gov.uk/government/publications/teaching-excellence-framework-subject-level-pilot-specification.

257. **Imbalance of metrics**: Where the metrics appeared not to be operating in a neutral way when applied in accordance with the step 1a initial hypothesis formula, two key areas were identified by the panel. The first of these related to a perceived overweighting of employability, and a second identified as a particular concern by the students related to the weighting of the NSS scores which seemed to disadvantage the student voice. The students commented:

> 'The half-weighting of the NSS, when taken with the formula at 1a, means that negative student views are deemed more important than positive views – poor NSS scores can put you into a Bronze at 1a, but no matter how positive your NSS scores are, they alone cannot result in a Gold. The disparity between these two is bizarre.'

258. **Small provision**: This included subjects which were reviewed with small student numbers or non-reportable metric data, so that it was difficult to make any confident judgement regarding the final outcome judgement. In Model B, this scenario included three cases and in two, a robust discussion was held by the panel about whether a rating could take place.

259. **Default to Silver in the absence of data**: The specification required a default to Silver where insufficient evidence in the metrics made it impossible to conclude any other judgement. The panel felt discomfort in defaulting to Silver without the presence of data, and especially so, when missing data could potentially have changed the hypothesis at 1a or 1b.

260. **Limited reference to the subject being assessed within the group written submission**: The panel found it difficult to achieve a holistic final judgement where the written submission for the group of subjects did not explicitly identify the subject being assessed.

261. In completing Model A, the panel expressed confidence in concluding ratings and identified that it found the longer subject-specific submission in Model A made it easier to conclude a more holistic view. Cases selected in Model A included fewer small providers, which meant fewer issues arose around small samples than was found in the Model B assessment. Less data in the metrics was non-reportable (this may have been an artefact of the pilot selection). The panel found that this, combined with the added detail provided for each subject in the written submission, meant robust ratings were reached even where subjects had smaller numbers of students.

262. A final 'wash-up' session in each model allowed the panel to review a small number of cases in which judgements were felt to have been of a lower confidence within the discussion and the panel agreed that this opportunity made it feel that its final decisions were robust. In two cases, where a robust conclusion became difficult to reach, the whole panel was requested to review the submission with a time delay (overnight) to allow time for review, before concluding a whole panel decision. The panel agreed that extra review by all members was helpful in enabling robust conclusions. Any provider ratings not included on the agenda due to higher consistency of agreement were added to the second day if time became available to assess them.

Comparison of the models

263. It is important to consider the comparative benefits of the two models and, to achieve this, exploring whether one model was able to conclude more or less robust assessments than the other was helpful. The Medicine and Health Sciences Panel observed that subject-level evidence in Model A written submissions allowed more holistic submissions and reduced the

possibility of the presentation being about a different subject (except in subjects allied to medicine). In Model A, it was less reliant upon the metrics, and made decisions even where providers were small or specialist. One academic panel member said that they found it easier to identify whether a subject was Gold, Silver or Bronze according to the criteria in Model A, and while the cases varied in their relative strength within these criteria, they felt less need to use borderline decisions. The panel suggested that the exercise was more balanced and it was able to use evidence with the criteria for Gold, Silver and Bronze more effectively to conclude robust decisions.

264. In Model B, the panel referred to the criteria frequently, but found difficulty when faced with conclusions deriving strongly from the metrics. One example was cited around the decision to be reached regarding 'all students' in the Gold criteria, which if taken literally could mean that no provider could achieve a Gold rating. This led to considerable discussion around the concept of 'all students' when the focus in Model B related to the split level data in the metrics and where one group with particular characteristics achieved negative flags or absolutes when the remaining population achieved positive ones. This was compounded where limited provider submissions did not articulate the needs of all groups. The panel felt that the criteria were not fit for purpose within Model B. Fitness for purpose of the criteria was also a feature in Model A but to a lesser extent.

Operating at scale

265. When considering scaling up these models, the panel observed the following conclusions.

Burden on the panel members

266. The panel felt that the longer submission for each subject in Model A was less work because it did not have to search the document for material on the subject. It did not foresee issues in using this style of submission on a larger scale.

Number of cases for review

267. In Model A, it was possible to review all cases in the whole panel, which assured a similar robustness in processes, whereas in Model B this was not possible due to the higher number of cases reviewed so some of these required sampling and moderation instead. This has implications if future work includes a Model B approach as consideration regarding moderation and sampling may be required.

Use of exceptions

268. The panel questioned whether the exception process in Model A was robust compared to including all subjects in Model B and asked whether universities could choose to ask to have subjects included in the process. It felt the ideal model would be not to have exclusions but for all subjects be considered, but recognised that this could be expensive as more panels would be required. This point was supported strongly by the students and new graduates in the panel who commented:

> 'From a specific student perspective Model B is preferable as it ensures all subjects have had oversight and consideration to present an actual outcome, rather than a risk-based algorithm based on metric flags'.

Support for providers

269. The panel found that some providers were advantaged in their capacity to deliver robust written submissions which narrated their situations well, and that others were disadvantaged by being unable to achieve this. The panel observed that the larger providers appeared more able to provide more comprehensive narratives, but commented that in both models some clear guidance for providers was needed regarding the narration of their activities to demonstrate excellent practice and the impact of this, as well as addressing the challenges they faced for development. If this was to be scaled up, then some greater preparation for some providers would be needed to enable them to provide improved written submissions.

Devolved nations

270. The panel reviewed a small number of submissions from Wales and Scotland. It found no evident differences between these submissions and those submitted from English providers. No special circumstances appeared to apply.

271. The panel felt that ratings achieved by the end of the individual and panel processes were produced in a robust manner with checks and balances in place to ensure full review. They identified that ratings in Model A were more robust in comparison with Model B and that larger providers had better evidence to support a movement out of the Silver rating, particularly in Model B, because the larger providers frequently had more reportable metrics. This disadvantaged small providers. The subsequent provision of a 'No rating' for these providers alleviated panel concerns in Model B where non-reportable metrics were combined with a limited written submission.

Quality of the evidence

272. Consideration of the quality of the evidence provided for review is critical. Specific points were observed.

Presentation and utility of core and supplementary metrics at subject-level aggregation

273. This material was found to be useful in determining the initial outcomes but was particularly helpful in Model B, where in some instances the metrics were the only available data which could be used make a decision relating at subject-level.

Employment data

274. A specific issue was raised around employment metrics within medicine, nursing and pharmacy. In these subjects, employment is currently expected to be high because of workforce shortages in these areas. Further, where part-time study exists, this provision is frequently for qualified professionals already working full time. The quality of the markers was weakened because benchmarks for the employment indicators were high and this challenged the validity of using high absolutes as flags. For example, it was found in one instance that the provider had achieved a double asterisk for a very high absolute value indicating that a positive flag could be considered at step 1b, but remained below the benchmark. The panel suggests that the use of high absolute value markers in employment for subjects where employment is expected to be high, should be considered with caution and that use of two full markers for employment is offering too much weight in this element.

Tariff data

275. While the tariff data was noted to be interesting it did not identify widening participation in an academic sense, as levels did not differentiate the type of pre-entry programmes studied. It was identified that it could be useful to consider benchmarked contextual data relating to the types of entry qualifications, e.g. Access and BTEC programmes, which could consider the provider's commitment to academic widening participation and the nature of likely learning gain of students from entry to completion.

Impact of CAH2 subject classification

276. Use of the CAH2 classification worked well in Model A for medicine and health sciences but less well where subjects were aggregated to a group submission in Model B. The panel had specific concerns about 'subjects allied to medicine not otherwise specified' (SAM) as this subject frequently included a wide-ranging selection of courses, some of which were included as part of joint honours programmes or at the request of the provider and bore little relation to other subjects within the group. Submissions were found to aggregate students who had very different experiences and outcomes and the panel questioned whether SAM could be treated equitably when individual subjects had no opportunity for a single subject submission even within Model A.

277. The panel reviewed subjects included within the CAH2 classification in a final session of the second panel meeting and concluded that:

- midwifery and nursing should be separated as they are separate professions
- SAM requires reconsideration to enable greater equity of assessment of subjects.

278. If CAH2 is to be used this should be made clear to future students when outcomes are reported to inform them about the ratings their chosen subject has been awarded. This is particularly important if the model selected ultimately includes a Model B type group submission, if subjects are awarded by non-exception as in Model A to the provider rating, or where the outcome is conferred on a group of subjects (e.g. in SAM). Students and new graduates identified this point as particularly important for transparency, noting: 'The use of CAH2 subject grouping would need to be made clear to students and applicants when communicating TEF outcomes'

Presentation and utility of contextual data

Office for National Statistics maps

279. The presentation of contextual data was found to add variable value to the evidence to be assessed. In the case of the contextual data for the institution, this was found to be helpful in both submissions and some members of the panel commented that it would also have been useful to have seen a summary of the provider submission also. The Office for National Statistics maps were found to have some use in the provision of context and aided discussion.

Teaching intensity data

280. This data was not found to be helpful. Sample sizes for students were too small to be representative and the panel was also unclear about what interpretation should be placed upon

the teaching intensity outcome provided. In a number of regulated courses within medicine and health sciences, the number of teaching and practice hours are professionally regulated and this could have influenced the outcomes and the reporting of this data.

Grade inflation data

281. Assessors found limited inference could be made if grades had changed. Discussion related to the idea that if teaching and improving outcomes for all had improved, the outputs should be better and this might include higher grades and classifications for graduates even if the standard and level of learning retained expected rigour at the level of study of an accredited course (e.g. bachelors' degree). It was therefore found to be interesting when a change in outcome grades had been narrated well in the submission as an impact, but the assessors identified that without clear parameters, little judgment could be made in respect of whether evidence added or detracted from the holistic judgement.

Provider approaches to the submission

282. In the Model B submission, some providers took the approach of separating out individual courses within the subject group, while others focused upon the elements of the TEF activity. Splitting into subjects where a small number of courses were included within a group made submissions easier for assessors, as they could identify which material related to each subject. Where the provider had focused upon the elements of the TEF to describe its achievements across the subject group, rather than identifying specific subjects, the submission responded to the required elements in a generic way, which made it difficult to identify whether the aspect being described was best practice in one subject rather than the same for all.

283. Lack of clarity also occurred in Model B where the limited number of pages made it difficult for all courses within a group to be illuminated fully and particularly where the provider delivered a number of subjects or where the submission included 'subjects allied to medicine'. Where many subjects were included, some were not mentioned in the group submission at all. The panel found it difficult to conclude a confident holistic decision in cases where the above factors combined. They concluded that group submissions in Model B offered potential for a 'halo' effect between different subjects. Finally, while recognising the need to make holistic judgements incorporating all submitted evidence, the panel discussed the balance around use of alternative evidence in the submission, including weighting to be given to internal surveys, single student or examiner comments, 'awards' or other evidence established for other purposes.

Quality and robustness of the assessment process

284. The Model B assessment was undertaken by the Medicine and Health Sciences Panel before undertaking Model A and this first exercise represented the application of new skills gained during its initial intensive training and calibration exercise. This is important as the panel had limited points of reference regarding its confidence in decisions. Nonetheless on each occasion, panel members had concluded judgements which they expressed with confidence before the review meeting and this was substantiated in panel feedback, albeit with recognition of the limitations of the process and their own abilities.

Does the three-step process transfer successfully to both models?

285. The panel identified the three-step process worked in both models, but some features could have been improved. During the panel individual assessment stage, where panel members submitted outcomes which required review by the TEF officer, queries mostly focused upon stage 1a, which was provided by a metric based calculation. The panel discussed whether it believed that it would be of greater benefit to have submissions pre-calculated at 1a, as this did not require subject expertise and a pre-determined 1a could offer greater consistency and reduce the need for the TEF officer to 'clean' this data. No conclusions were drawn. Some thought that there was value in working out the calculation, whilst others supported the stage being predetermined. All identified the initial hypothesis as an important marker for continuing to assess the submission.

286. When working on Model A, the panel identified at step 1b, a need for greater clarity about the way in which absolute values which were not statistically significant were treated, as the panel interpreted the process in two ways, which it felt changed their thinking. For instance, such absolute values could either:

- Be interpreted as flags and added to the core metric flags at 1a prior to considering more holistically the full set of core metrics data to determine the initial hypothesis at 1b
- Not treated like core metric flags, but considered holistically alongside the full set of core metrics data to determine the initial hypothesis at 1b.

287. The panel felt this subtle distinction was important, as if it took the former it felt it would be justifying a 'Gold' starting point, whereas in the latter case it would be Silver, although it did agree that in the final holistic judgment the inclusion of other factors may mediate differences. The panel also observed that the written submissions had significant influence upon the outcome and noted that it could have benefited on having a greater focus on evaluating these provided in the initial training. In respect of the TEF Gold, Silver and Bronze criteria, after using Model B the panel commented these should be reviewed to make them more applicable to the evidence reviewed. This was less notable in Model A, where submissions were more specific, but the panel still would like to see review of the criteria. It also observed that it was difficult for smaller providers to move between bands due to low confidence in small numbers.

Availability of information between subject panel and main panel

288. Information summarising key points from the main panel was offered by the panel chair and deputy at the commencement of each panel. Information in between panels was circulated by the TEF Officer. This worked, although pace of change during the pilot meant that advice to panel members changed. This was recognised as unavoidable in the pilot process.

Coverage of panel expertise and impact of different panel member roles

289. Ways of working were discussed amongst the Medicine and Health Science Panel in a collegiate manner at the beginning of the exercise and space was allowed for further discussion and feedback throughout. This allowed everyone to participate equally in decision-making and for consensus to be achieved. The panel was a respectful and caring group of professionals who attended to the views of others and quickly worked well together to achieve the work required. A mixture of more experienced team members who had undertaken TEF Year Two assessment worked extremely well in supporting those new to the discussions and in

developing the ability of all in asking searching and detailed questions. It was agreed that in each case reviewed, the initial presentation would be made by the TEF officer presenting core details of the subject to be reviewed and then panel assessors allocated to the subject case would present their findings in greater detail. A wider critical discussion including the whole panel addressed points of concern and considered the final holistic judgment to be made. The panels were conducted by the chair (an academic) and deputy chair (a new graduate) and managed by the TEF officer. The TEF administrator was in attendance to offer advice on technical issues and record technical details.

290. In the feedback sessions, the panel commended the ways of working and noted the following points.

Panel meeting organisation

291. The panel noted that they felt that meetings had been well led by the chair and that the team (including the OfS representative, chair, deputy chair and TEF officer) worked together well, bringing skills and knowledge together to support them. One academic summarised: 'The panel has worked together incredibly well.'

Student involvement

292. Students and recent graduates were active panel members, whose contribution was integral to the discussions at all times. They noted that there was an imbalance in numbers between academics on the panel and others, including themselves. The students commented:

> 'Though student members were considered equal members of the panel (and acted as such) the ratio between students and academics would implicitly mean the specific student voice was lessened (quite often 3:1 in panels).'

Subject-related expertise

The role of the panel and its constituent members needs to be considered carefully. The panel subjects with experts 'in the room' may receive different levels of scrutiny to those without. This could also apply similarly when a provider moved a subject to a different group. A strong need for subject expertise emerged in Model A due to the more detailed information which was included within the longer subject related submissions. In Model A, each panel assessing a submission included someone with some subject expertise (where this was available within the panel). It was noted that for some subjects this was not the case (e.g. dentistry) and it was recommended that major subject groups must include an expert on the panel. The panel highlighted a challenge in finding experts without conflicts in small specialist areas and noted the number of experts required to cover all regulated subjects if Model A was scaled up.

Design considerations

293. The panel proposes that, in both models, consideration of the relative weight of the metrics used is essential and improving the Gold, Silver and Bronze criteria would enhance clarity for providers in respect of decisions made. Key considerations for each model are listed below.

a. Model A design considerations:

i. The longer subject-specific submission should be continued but some consideration should be given to subjects which had included within them a large number of subject groups (e.g. SAM). Consideration should be given to determine how this subject can be assessed equitably and how useful information could be offered to students attending one of the courses.

ii. The panel noted that the process of including exceptions in Model A would have meant that any exceptions whose rating would be changed by very high or low absolute values would not have been included. It expressed concern about exceptions, particularly when these were assessed only on the metrics of the provider.

iii. The panel appreciated the presentation on the ratings offered by the TEF team at the end of the process in Model A and felt that comparisons made enabled it to moderate its decisions and conclude more confidently.

b. Model B design considerations:

i. The group submission should be reconsidered.

ii. In subjects allied to medicine not otherwise specified, writers of submissions need more guidance on what to write to capture individual subjects.

iii. Where subjects are grouped by provider it should be recognised that panels may not have the expertise to make robust subject-specific judgments and resources may need to be made available. The proposal to review other subjects within a group if data was non-reportable was not found useful in the experience of this panel.

iv. Absolutes are at provider-level and are expected to be high in some areas. If the subject is moved to another panel this could have an undue influence.

v. Supplementary metrics including teaching intensity and grade inflation should be reconsidered and, if included, greater guidance given regarding evaluation.

Subject-specific considerations

294. The panel raised subject-specific issues related to the weighting of employment data and the influence on the process in some subjects and this is addressed in the section on quality of the evidence. A new consideration related to ‘continuation’ where 100 per cent continuation might suggest that students were being retained even if professionally unsuitable. Reflection of this data being nearer to subject benchmark may be important but this also impacts upon the validity of the high absolute in these case. Students and new graduates commented:

> ‘Employment in the majority of the medical subjects has very little weighting on the applicants’ decision – generally graduates from these courses can find employment without hassle. (Also, it can be six or more years from application to working the profession, meaning the DHLE statistics are broadly irrelevant)’

A key feature for the whole panel also related to the half-weighting of NSS results against the two whole markers for employment.

Segmented panel member views

295. The perspective that everyone was equal round the table was much valued by all members of the panel. Included below are specific perspectives relating to panel member views:

Student representatives

296. It was identified that the panel included both students and new graduates with different experiences, and that the new graduate voice was extremely valid and it was proposed that if the two could be identified separately, this would be a more accurate reflection.

Professional, statutory and regulatory bodies and employers

297. PSRBs' and employers' views related strongly to ensuring the professional credibility of the TEF and of their own professions. Where evidence was omitted from the written submission which assessors felt was important in reaching a credible TEF decision from a professional perspective, this presented challenges.

298. A case example shared by PSRB representatives illuminates this point. The pharmacy registration examination takes place one year after students graduate. The examination offers a public ranking of universities' success compared to others. The credibility of TEF may be at risk if the subject achieved 'Gold' but was identified at the same time to be a poor performing school in these professional rankings. Employer and PRSB representatives commented:

> 'Employability of the student is not currently measured in TEF in terms of "fitness to practice" beyond graduation. Consistency between TEF and existing post-graduation measures of fitness for practice should be explored'.

299. The panel held a robust discussion about the relative merits of external evidence and recognised that the pharmacy registration examination and the TEF were measuring different outcomes currently. They acknowledged the importance of further discussion and highlighted that TEF must be professionally credible, if it is to be useful to students in future.

Widening participation

300. The value of including employers on the panel actively was noted by both those participating and other colleagues. They proposed that this should extend to the widening participation experts in order to ensure a more meaningful contribution.

Feedback on submissions

301. The panel recommended that clear guidance and training be offered to providers to enable them to articulate effective written submissions. Whilst there was evidence of very good submissions across the sector there were a number of poor submissions also. In the best submissions, assessors found the following:

a. Providers demonstrated strong impact of their actions to enhance provision.

b. Providers included narrative for programmes where data was non-reportable or missing.

c. Providers reflected on the TEF metrics in both positive and negative instances to celebrate excellence and address challenges.

d. Providers included material which was different from the evidence in the metric workbook and this was narrated in the context of the metric.

e. Providers illuminated their consideration of 'all' 'students in respect of the measures of excellence. While some did focus on specific groups of students (i.e., the majority group), the best addressed the needs of other groups.

f. Providers' narratives accurately reflected the outcomes identified in the metric workbook, when they were citing metric outcomes.

g. Care was given in expressing the representativeness of evidence included within the written submission. The panel noted the student and new graduate comment:

'There was a very varied approach to inclusion of students and student representative bodies. Students felt frustrated that the specification encourages student participation in the process but does not hinder institutions that do not facilitate this. The student voice should have a significant impact here and may provide insights beyond high-level statistical or budget statements'.

Potential impacts

Protected groups, widening participation considerations

302. A concern was identified by the panel in relation to ensuring all groups receive excellent education and this is summarised well by the students who participated:

'With the current sector focus (rightfully) exploring the BME attainment gap and the Gold category encompassing good outcomes for "all", student members felt particularly uncomfortable awarding a Gold rating for organisations with a negative flag in the BME split metrics.'

Aspects of quality and student experience and applicant choice

303. In subjects allied to medicine not otherwise specified, distinct programmes are not identified to applicants.

Discipline-specific impacts

304. As noted in this report, it is important to ensure that the credibility of TEF and the professions is maintained in order to continue to enjoy public confidence.

Wider sector impacts

305. The impact of TEF decisions on applicant choice is considered critical and it is critical that the consumers of this exercise are not misled by TEF. In order to ensure robust and informed decisions (critical in ensuring a credible impact of TEF and minimising reputational risk), a longer time period for development of the tools, calibration, training and assessment review for assessors should be explored in order to ensure greater consistency in the process and outcomes. This may extend to the development of a pool of experts for whom a lesser form of updating will be possible.

306. Providers in the UK already receive threshold institutional accreditation to ensure that they are fit to offer provision. It is important the TEF, which focuses on excellence, does not give out

the message that provision is inadequate when this is not the case, thus leading to a lack of confidence in the sector which would have serious impacts on professional workforce. The student members commented that:

> 'Bronze is implicitly viewed as negative, when in fact it is supposed to represent above the baseline – in awarding a Bronze, students often felt that applicants may see this provider as under-delivering or under-performing when in fact that is not the case'.

Additional observations

307. The panel felt the template for feedback offered a good opportunity to reflect on the process and evaluate models comprehensively. Additional observations related to the very short timescales and the time of year in which the pilot took place (spring) which included a number of deadlines falling close to bank holidays. The chair and deputy chair recognised that the half-weighting of cases in both main panel and subject panel was extremely helpful and facilitated their contribution. Finally, panel members previously involved in assessing TEF Year Two identified that the used of 'Trios' in that assessment was helpful and while not used in this pilot due to the approach suggested they might be considered as a useful inclusion in future approaches.

Conclusion

308. This report has illuminated evaluative observations and critical commentary of the medicine and health sciences pilot subject panel during its review of two models (A and B) during the subject pilot review exercise in the spring of 2018. The panel recognises that supporting the evaluation of teaching excellence for future learners wishing to access professional and vocational programmes in the subject area of medicine and health sciences offers the benefit of better informed future healthcare professionals who will be able to choose excellent teaching programmes which meet their needs. These findings are offered in the spirit of constructive and collegial support for the enhancement of teaching excellence by the panel, which would like to commend the continuing efforts of the TEF development towards an effective future model of teaching excellence.

Report author: Prof Carol Hall, Chair of the Medical and Health Sciences Panel

Table 5: Medical and Health Sciences Panel members

Chair	
Prof Carol Hall	Director of Undergraduate Education, School of Health Sciences, University of Nottingham
Deputy Chair	
Aaron Lowman	Former Students' Union Vice-President, Brunel University London

Panel members	
Dr Tracey Cockerton	Deputy Dean, Middlesex University
Prof Lesley-Jane Eales-Reynolds	Director, Institute for Innovation in Learning and Teaching, University of West London
Charlotte Freitag	Former Class representative for Psychology, English Literature and Sociology, University of Glasgow
Dr Sue Haines	Assistant Director of Nursing (Professional Development, Education and Workforce), Institute for Nursing and Midwifery Care Excellence, Nottingham University Hospitals NHS Trust
Christopher John	Head of Workforce Development, the Royal Pharmaceutical Society
Prof Neil Johnson	Faculty Dean, University of Lancaster
Dr Penny Joyce	Associate Head Education, University of Portsmouth
Alykhan Kassam	Former Pharmacy and Faculty of Life Sciences Student Representative, University of Bradford
Prof Julie Mcleod	Pro Vice-Chancellor, Oxford Brookes University
Leonie Milliner	Chief Executive, Association for Nutrition
Rory Murray	Former Students' Union President, University of Kent
Prof Robert Norman	College Academic Director, Leicester Medical School, University of Leicester
Dr Jan Quallington	Head of Institute, Health and Society, University of Worcester
Prof Susan Rhind	Assistant Principal and Director of Veterinary Teaching, University of Edinburgh
Prof Trudie Roberts	Director of the Leeds Institute of Medical Education, University of Leeds
Eleanor Rosario	Student Representative for Medical Society, Faculty Learning and Teaching Committee and Peer Teaching Society, University of Sheffield
Prof Richard Tong	Assistant Principal (Higher Education), NPTC Group of Colleges
Philippa Tostevin	Head of the Centre for Clinical Education, Course Director for Medicine, Reader in Surgical Education, Consultant Ear Nose and Throat Surgeon, St George's, University of London

QAA TEF officer: Kevin Kendall

Natural Sciences Panel report

Executive summary

309. This report expresses the views of the Natural Sciences Panel about the operation and outcomes of the TEF subject-level pilot 2017-18. The panel considered that it added value to the assessment process. We recommend that the balance between metric weightings and qualitative, and ultimately holistic, judgements, be reviewed to give equivalence to the latter. The panel valued the diversity of its membership, in terms of student, staff and employer input and gender balance. We would encourage a greater diversity of applicant going forward.

310. Of the two models we were asked to operate, some elements of both worked well, but neither is currently ready for full roll-out. We were particularly concerned at the assumptions needed to operate Model A and its inability to guarantee a robust outcome for subjects which were not exceptions.

311. Reports at a granular, subject-level (e.g. mathematics) had an authenticity and clarity that facilitated judgement, and were clearly tools for enhancement. However, it seems inequitable and risky to have only exceptional subjects submitted to the panel. Grouped subjects (e.g. natural sciences) were natural groupings in a minority of providers and offered less scope for clarity and a description of excellence.

312. Going forward we suggest the following, in addition to the central observations made in the previous paragraph:

a. The holistic descriptors need development to work at the different levels of an organisation to which they are applied.

b. A minimum size of cohort should be established for the provision of a rating. Subject-level metrics often make split data very partial and consideration should be given to the provision of some metrics for subjects and others at provider-level, where more nuance is visible due to higher student numbers in minority groups.

c. Consideration should be given to the award of a ‘provisional’ rating for new provision, to accommodate the need of institutions to diversify their provision, but recognising the partial nature of data on new programmes for a number of years after they begin to recruit.

d. Consideration should be given to a ‘private’ feedback document for the provider, which might give expert (i.e. panel) insight into the actions being taken by the provider to mitigate poor ratings or permit enhancement. We feel this offers a powerful enhancement tool for the sector and makes best use of an expert panel.

Were robust ratings produced?

313. Ratings were produced robustly according to the processes that were set out for us. There are several ways in which they may have been robust but not correct. Some of these are picked up later in this report. The key issues were that aggregation of subject areas gives artificial groupings and too little space to describe a subject group sufficiently. Where subjects were reported by exception, the standard and clarity of the reports were much higher, and in general the subject aggregations were sensible. However, there are clear flaws in the ‘by

exception' model that mean that the panel felt this lacked equity and accuracy. To take one example, a provider with an initial hypothesis of Silver could be raised to Gold, but a Silver exception, which had better metrics than the provider as a whole, could be retained at Silver.

314. The focus on metrics initially, especially the generation of an initial hypothesis, and a training that focused on dealing with metrics, meant that it was difficult to weigh the written submissions in any way other than against the metrics. A more nuanced approach would be possible, where metrics and text were weighed but not judged until a later stage.

315. The criteria on which the reports were based could be enhanced to provide guidance for both authors and auditors. The holistic descriptions don't work at subject-level. In addition, it might be that one description cannot cover the sector – some universities have a unique mission, and can only be judged relative to the degree to which they achieve that set of goals.

Comparison of the models

316. Some of this is covered in paragraphs 326 and 339. In addition, we feel there should be significant considerations of which model offers providers greater scope for enhancement. Despite the burden on providers, it is only by submitting a TEF report for each area of provision that there is an equal opportunity for subject areas across the provider to reflect on their practice and improve. Equally, should the TEF outcomes be of benefit to students in making their choice of provider, it is only by having a 'live' TEF rating for each area of provision that this remains a valid indicator. Otherwise, a poorly performing area with metrics in the same range as the provider as a whole will reflect the quality of provision overall.

Quality of the evidence

Metrics

317. We were constantly reminded in our deliberations that the metrics are proxies for only part of the criteria of excellence that define the TEF, and in some respects they are poor proxies. This is not to question their value, but to ask questions about their primacy in the judging process at subject-level. It is beguiling to look at the numbers first, but not necessarily useful. Although few of our judgements differed from the numerically generated hypothesis, we would report that this is because we followed the guidance given – back to the issue that our outcomes were robust but not necessarily reflective. We feel that NSS data is as useful as any other and should be weighted to reflect this. We found very little value in LEO data, which is historical and partial. Our use of peripheral data, such as maps or POLAR data, improved over the period of our work, but a great deal of the data we were using was of limited utility and should be considered in the balance of use versus effort to collect and to interpret well. We would welcome a metric that provided a more nuanced measure of value added, both in terms of a starting point as well as an end point, and in terms of a wider definition of value than a financial one. We are concerned at the opportunity for institutions to 'game' continuation figures to the potential detriment of students who are engaged in study for which they are not suited.

Scale

318. There are two key points here. The first is that the absence of data is disconcerting, but would be less so if there were more emphasis in the process on the written submission. The second is that a range of factors drive small but reportable data sets towards a step 1a initial hypothesis of Silver.

Submissions

319. Here we would highlight that longer, subject-based submissions were much more useful, and although more time-consuming to produce allow enhancement to pervade an institution. Some of the submissions in Model B were poorly written or poorly planned, with little use of internal data and little evidence of self-reflection. We wonder if this is because writing to an artificial grouping of subjects is practically difficult to do, and actually difficult to do well. It at least needs a longer word limit. A structure for submissions would offer clarity for authors and assessors, if a balance could be struck that ensured that this was an enabling framework within which diversity could be described and celebrated.

Context

320. It would be very useful to have a short written context for the submission, which might include data on programmes within a cluster subject, numbers of students on each course, proportion of shared provision with other disciplines outside this aggregation in the subject hierarchy, etc.

Quality and robustness of the assessment process

321. As noted above (paragraph 319), we considered that the primary use of metrics to define a hypothesis placed an undue weighting on the metrics and made the submission more of a mitigation than a primary source of information.

322. The panel worked well, with students, academics and employer representatives working seamlessly. The input of a WP specialist was very useful, as was the insight provided by the OfS and academic main panel members moving between panels.

323. The Natural Sciences Panel would have welcomed access to the provider-level submissions in Model A.

324. OfS guidance might usefully refocus in future to provide a balance of training on the metrics and written submission. However, the training process was very effective in allowing all panel members to start their work from a position of clarity and knowledge about the process.

325. Most outcomes that we assessed were initially Silver and remained so. A question was raised in our panel about whether there are enough categories in a threefold outcome. There was also some discussion around whether this tight clustering could be addressed by revised holistic descriptors.

Model A design considerations

326. The step 1a initial hypothesis is usually ‘correct’. Only three (of 20) submissions for which the panel made a judgement changed from initial hypothesis to final rating (in each case from Silver to Gold). However, we feel that this reflects our close following of guidance which made initial hypotheses weighty, rather than any sense that our deliberations necessarily confirmed the accuracy of the metrics in approximating the holistic judgment accurately.

327. The panel overall found the Model A submission to be more focused and more useful, with good submissions containing:

- evidence of self-knowledge and self-awareness

- evidence at subject-level as well as provider context
- evidence of impact of interventions
- acknowledgement of weakness as well as awareness of strengths
- understanding of issues arising in the metrics (and addressing of these issues)
- evidence of student voice contributing to the process (authentically, not just individual comments)
- reference to the assessment criteria
- context of the provider's subject provision (in terms of what the programmes were, and the numbers of students on them).

328. The panel had a high proportion of subjects with 'No rating', declining to rate seven (of 27) submissions, where it was felt that the provision was unsuited to the process. Six were Silver at initial hypothesis, one was Bronze, and the reasons for refusal to rate were:

- very small numbers
- submission was a disparate collection of programmes
- new provision
- nature of provision was unclear
- provision was a foundation year.

329. It is hard for small provision to get a rating, and it was suggested that a minimum headcount may form part of the criteria, along with a requirement for reportable data in all metrics. For small and new provision, it may also be appropriate to permit re-application after a shorter time period. Where metrics in future could be proxied or aggregated, the panel showed a clear preference for taking the provider rating, which would be more meaningful than taking subject data over a longer period of time.

330. It was noted that the process is a mix of norm and criterion referencing, which are sometimes in tension, and can lead to confusion or anomalies:

a. The Silver descriptor is 'excellent outcomes', which implies criterion-referencing. The Gold descriptor is 'outstanding outcomes' which implies norm-referencing.

b. An extreme example is one that dominates its benchmarks so it could not be 'outstanding' on its metrics.

c. The panel has gained confidence in making holistic judgements informed by submissions and metrics, especially when it began to recognise that for atypical providers excellence had to be based on assessment against self-defined missions and aims, rather than being required to exactly fit the current descriptors.

331. For many submissions, the panel felt that a mechanism for giving specific and confidential feedback to providers may help to support enhancement, and in some cases, particularly for new provision, or where significant development has been undertaken but has not yet come to fruition in the metrics, may affirm for providers that they are doing all the right things.

332. A student body with evidence of multiple and interacting disadvantage needs particular consideration, and reiterates the need for TEF descriptors to recognise institutions with large proportions of disadvantaged students who may be excellent or outstanding in their particular missions and to ensure that discussion around benchmarking is sufficiently nuanced and support by WP input at subject panel level.

333. Model A can generate inequity between subjects based on metrics at the same provider through exceptions, and it was felt that review of submissions by the panel was key to ensuring that holistic judgements were sufficiently nuanced to capture areas for recognition and development.

Model B design considerations

334. The step 1a initial hypothesis is usually 'correct'. Only four (of 50) submissions for which the panel made a judgement changed from initial hypothesis to final rating (three from Silver to Gold, one from Silver to Bronze), but see reflections on Model A for why we think this was the case (paragraph 331).

335. In reviewing the subject group submissions, confidence in our ratings was increased by evidence of impact, by institutional self-awareness, and by reflective practice (as in Model A). The panel identified the following deficits in Model B submissions:

- lack of evidence of impact (TEF Subject-Level Guidance has numerous examples of types of evidence, but many submissions are not making use of this guidance; perhaps condense this guidance?)
- not addressing the metrics
- holistic submissions are of less value than those aimed to each subject.

336. Types of evidence which are useful at subject-level include:

- positive PSRB comments help to give confidence (but their value is subject-dependent)
- external examiner statements have weight when taken with caution
- student views are useful if collected by robust and inclusive processes
- National Teaching Fellows are useful as an excellence measure for providers outside Scotland
- internal teaching awards may be useful, but processes need contextualisation.

337. As with Model A, development or enhancement undertaken by the provider but which lacked evidence of impact due to time since implementation was insufficient to increase a rating. In addition, evidence of evaluation processes in place is helpful even if it is too soon for evaluations to have been done.

338. For small programmes, it is hard to get significant flags, and as with Model A, we suggest that a threshold cohort size to be eligible for a rating would be appropriate. For new programmes, submissions ought to address the newness and any relationships with other programmes.

339. Overall, we felt that the page limit does not allow providers to fully address their subject provision, and is too tight to allow a case to be made for moving the rating up from the metrics. It is also more difficult to identify areas of good practice, and areas of concern in these submissions. The limited space means that the holistic descriptors do not work at subject-level as providers can't adequately address all criteria.

Segmented panel member views

340. We found this question difficult because by the time we asked it of ourselves, we felt more like a coherent group of experts. However, a key point to emerge at this stage of our discussions was the importance of hearing an authentic and pervasive student voice in the submissions. This was easier to do in the submissions we saw in Model A, where the word length was greater and the subject under consideration more coherent.

341. We also felt that a panel is well served by seeing submissions across a range of subject areas, and our PSRB experts felt that this was a strength and avoided any possibility of becoming blinkered or being too influenced by prior knowledge.

Feedback on submissions

342. We felt that we could identify good practice in submissions as detailed below, and that this guidance could be of use to authors in future. For us, a good submission showed the following:

- evidence of self-reflection
- evidence at subject-level as well as provider context
- evidence of impact of interventions
- acknowledgement of weakness as well as awareness of strengths
- addressing the metrics
- evidence of student voice contributing to the process
- referring to the assessment criteria
- giving context (in terms of what the programmes were, and the numbers of students on them)

Potential impacts

343. The most important risk that we identified was the possibility of TEF outcomes stifling innovation, with new courses taking time to become sufficiently established to be rateable. A provisional rating, based on the ratings of cognate subjects in the provider, might alleviate this risk. But there is also the risk that an innovation within a degree programme might lead to a

temporary fall in student satisfaction, and that institutions may become more risk-averse in future.

344. We are also concerned about an institution's response to a poor TEF rating. We saw several examples of a course doing much to improve, but not yet having demonstrable impacts – should the provider change the plan of improvement, or indeed close a poorly performing course, then the TEF would have served to work against the enhancement that is in its core intentions.

345. Finally, a provider might exhibit more complacency towards a discipline that is performing within average metrics, especially in Model A, and we would suggest that whatever model is ultimately delivered incentivises enhancement and permits an institution to focus on all disciplines, not just those that are atypical in metrics distribution relative to the provider as a whole.

Additional observations

346. We felt that an expert panel is key to good decision making, and that the training and experience of undertaking the TEF pilot assessments was very useful in building the expertise base within the sector.

Report author: Prof Sue Rigby, Chair of the Natural Sciences Panel

Table 6: Natural Sciences Panel members

Chair	
Prof Sue Rigby	Vice-Chancellor, Bath Spa University
Deputy Chair	
Martha Longdon	Students' Union President, Nottingham Trent University
Panel members	
Prof Mark Clements	University Director of Learning, Teaching and Student Experience, University of Middlesex
Prof David Coates	Professor of Life Sciences, Biological and Biomedical Sciences Education, University of Dundee
Michael Cottam	Assistant Principal Higher Education, Myerscough College
Eleanor Furness	Institute of Biological Environmental and Rural Sciences Representative, Aberystwyth University
Dr Abigail Hind	Director of Academic Services and Academic Registrar, Harper Adams University

Matthew Nishaan Kenyon	Medical Genetics and Genetics Course Representative, Queen Mary University of London
Prof Duncan Lawson	Pro Vice-Chancellor (Formative Education), Newman University
Lauren Marks	Former Students' Union President, Education Officer, Institute and Course Representative, Aberystwyth University
Dr Mary McAlinden	Head of Department of Mathematical Sciences, University of Greenwich
Dr James McEvoy	Senior Lecturer and Associate Dean (Education), Royal Holloway, University of London
Nicole Morgan	Education Policy Manager, Royal Society of Chemistry
Ali Orr	National Centre for Universities and Business, Talent and Employability Consultant
Prof Katharine Reid	Deputy Head of the School of Chemistry, University of Nottingham
Prof Richard Thompson	Professor of Experimental Physics, Imperial College London
Harry Williams	Former Course and Faculty Representative for Natural Sciences, University of Keele

QAA TEF officer: Stephen Ryrie

Social Sciences Panel report

Executive summary

Introduction

348. The TEF Subject Pilot Social Sciences Panel comprised 13 academics, six student members and one employer. The panel met on 11 and 12 April to produce subject-level ratings under Model B and on 10 and 11 May for Model A. The panel was chaired by Prof Neil Ward, Deputy Vice Chancellor of the University of East Anglia, with Diarmuid Cowan, the Students' Union President at Heriot-Watt University, as deputy chair.

Robust ratings

349. The panel was generally confident that robust ratings were produced. However, this confidence comes with one important caveat. Where submissions involved smaller numbers of students, the panel had serious concerns about the robustness of the ratings that could be produced (see paragraphs 355 and 365). It is encouraging to note that where ratings were separately produced for the same subject areas under both models the same ratings were arrived at in every case.

Non-reportable metrics

350. There were serious concerns about the robustness of ratings when student numbers were smaller and there were non-reportable metrics. The panel felt that when the data was insufficient, the more robust, statistically informed and methodical approach to assessment was seriously undermined. A considerable proportion of time in the Model B assessment meeting was spent on smaller submissions and grappling with the implications of limited data. The panel judged that there needs to be at least 30 full-time equivalent students in the reportable metrics for ratings to be robust. The Office for Students' TEF team should carry out its own quantitative analysis to inform the development of a minimum size threshold for submissions. (It is quite possible that this should be higher than 30 students, but the panel would expect it not to be lower).

Comparison of the models

351. The panel found the assessment process much more straightforward for Model A because the written submission wholly mapped onto the metrics workbook. Under Model B, assessors had to search through subject group submissions to identify material relating to the subject area under consideration, which was awkward and time-consuming. (This difficulty in Model B could be mitigated if ratings were given at the subject group level rather than the subject area level, or if providers were required to structure their subject group written submission in such a way that individual subject areas were presented in discrete subsections). There was a concern among some panel members about the way exceptions are generated which was felt to weaken the purchase of Model A (although other panel members were relatively more comfortable with the exceptions approach in Model A).

Quality of the evidence

352. The panel was generally comfortable with the core and split metrics data, except when submissions involved small numbers of students and suffered from non-reportable metrics. The panel did not make much use of the Office for National Statistics data maps. There were

concerns expressed about the quality of the employability (employment and highly skilled employment) data in the core metrics and the long lag times in the LEO-based supplementary metrics. There was a particular concern that the employability metrics did not include data for international students studying in the subject areas. Notably, these concerns were particularly expressed by the panel's employer representative. The quality of evidence presented in written submissions was variable. There was some concern that several further education colleges seemed less able to marshal convincing additional evidence in their written submissions.

Quality of the assessment process

353. The panel members became more comfortable with the assessment process over the two stages of the pilot exercise. There was some concern about the ways the initial hypothesis could exert significant influence throughout the assessment process – an 'anchoring effect.' However, the panel became increasingly confident in coming to holistic judgements in the round, and referring to the ratings descriptors. The panel did feel, however, that the ratings descriptors require further development for use at the subject-level and that the Gold descriptor, in particular, set the bar too high.

Model A and Model B design considerations

354. There were no particular issues around the submissions (metrics workbooks and written submissions) under Model A. There was some discussion around the approach to exceptions in Model A, the panel recognising that only a proportion of providers' subject areas were actually being fully assessed. Panel members felt it would be useful to know what courses are contained within the subject area under consideration under both models. For Model B, the written submissions were very difficult to navigate and it would have been much more preferable to have had dedicated sections within each subject group submission that dealt with the individual subject areas.

Subject-specific considerations

355. It was noted that there was not as much subject-specific discussion as may have been expected. Architecture 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 more explicit guidance and requirements for providers on how to present information about PSRB accreditations.

Segmented panel member views

356. Student members played a full part in the deliberations and the contributions made by student and academic members were treated equitably. Student members felt that, overall, employability data was overemphasised in the assessment process and would have preferred more emphasis on WP data. Student members suggested that it would be helpful to have an explicit step in the assessment process to consider WP and student voice issues.

357. The panel was able to discuss PSRB issues, but a PSRB representative might have strengthened the panel's treatment of these issues. The employer panel member played a full role in the discussions and brought particular concerns to bear around the quality of employability data.

Feedback on submissions

358. Providers should be given more explicit guidance on the following:

- addressing evidence gaps where there are non-reportable metrics
- evidencing impact of performance improvement initiatives
- actions in response to actively engaging with students and their representatives
- the significance of PSRB accreditations.

Potential impacts

359. There was a concern among the student panel members, which was widely shared, that the focus on outcomes split by WP characteristics could potentially provide a disincentive for providers to recruit from underrepresented groups. It was felt that evidence of satisfaction, attainment and outcomes among WP groups therefore needs to be considered alongside the contextual information on student intake, and any strategies to strengthen representation of under-represented groups in the providers' student bodies.

Additional observations

360. The panel operated well and there was a good range of expertise and highly informed discussion to arrive at ratings. The sessions covering feedback on the assessment process were particularly productive, and there was a high degree of consensus around the majority of the insights generated from the exercise.

The robustness of ratings

Small submissions and non-reportable metrics

361. The panel had serious concerns about those submissions with smaller numbers of students in their metrics data. These concerns were most acute when there were less than 30 full-time equivalent students. Small numbers of students in the data made assessment based on metrics problematic and the panel felt that this problem was serious enough to risk compromising the legitimacy of the TEF process. When the data is insufficient, the robust, statistically informed and methodical approach to assessment is seriously undermined. The panel judged that there needs to be at least 30 students in the reportable metrics for the ratings to even begin to be considered robust, and possibly more. For larger cohort submissions, under both Model A and B, there was a reasonable level of confidence among panel members in the robustness of the ratings. Metrics were important in informing the assessment process, but it was the holistic mix of core metrics, supplementary metrics and the written submission that enabled panel members to come to a carefully considered judgement.

The ratings descriptors at subject-level

362. There was some unease at the seemingly high standard expressed in the Gold rating descriptor when applied at the subject-level. The rating descriptors will need further work to be adapted for subject-level TEF. The panel was anticipating that the general distribution between Gold, Silver and Bronze in our subject-level exercise would be broadly comparable to TEF Year Two. However, under each model, we saw an underrepresentation of Golds in the social sciences.

The ratings scale

363. There was some interest in the potential advantages of a five-point rating scale. This could, for example, be introduced through having ‘Starred Bronze’ and ‘Starred Silver’ for the strongest submissions in those two categories. It was felt that a five-point rating system would provide stronger incentives for providers to strive to improve their performance and their ratings over successive assessment exercises than a three-point rating system.

Comparison of the models

364. The panel felt that the structure of the written submissions was more straightforward and easy to use in Model A compared to Model B. This meant that there was more detail provided, a stronger basis for discussion and decisions, and the panel was more confident about the decisions arrived at. Although the differences in the models will have been mitigated by the OfS’s decision to allow panels to make a ‘No rating’ judgement, overall panel members felt that decisions were more robust under Model A than Model B because of the five-page written submission.

365. The written submissions for the subject groupings under Model B did not work well and assessing submissions under Model B was significantly more burdensome for assessors than Model A. This was principally due to the difficulties in navigating the material contained in the subject group written submissions.

366. This was a particular problem where metrics were based on small cohort numbers and so greater reliance had to be placed on the written submissions. Few providers were able to produce written submissions that genuinely helped the assessment process under Model B. Providers tended to synthesise material across subject areas within a subject group, but because the rating was produced for the subject area level, assessors had to scan through written submissions and search for material relevant to the particular subject area under consideration. There was a concern that the uneven coverage in many Model B written submissions where there were multiple subjects may lead to a greater focus on metrics in reaching a judgement.

367. It would have been much more straightforward for the assessment process if providers had been guided to subdivide their subject grouping written submissions into the component subject areas. Better still would be a separate written submission for each subject area. In contrast, the opportunity to provide a single subject focus in the written submission in Model A enabled a good balance between metrics, written submission and descriptor in reaching a judgement.

368. At the subject panel, there was no interaction between the provider-level submission and the subject group submission. Some providers’ subject group submissions made reference to the provider-level submission, but the panel members did not see or consider these. Under both models, panel members would have preferred to have seen more institutional context around the subject area being considered.

369. In general, some further education colleges seemed particularly hampered by small cohort numbers which compromised the assessment process. Moreover, several colleges struggled to produce compelling additional data and evidence in their written submission, compounding the poor coverage of their metrics. There did not seem to be significant differences between the

models in their handling of the range of providers. Five pages of dedicated writing for Model A Subject Areas gave sufficient space to discuss any issues particular to a provider type.

Quality of the evidence

370. The core metrics were used extensively and systematically, in accordance with the guidance on the assessment process set out in the TEF specification document. Supplementary metrics were referred to, but the experience was mixed and sometimes they appeared to contradict core metrics. Generally, less weight was placed upon the supplementary metrics than the core metrics. Supplementary metrics were most helpful when considering those institutions with a distinctive mission around disadvantage and social mobility. Very little use was made in the discussions of the Office for National Statistics data maps. It was only very occasionally that they were useful. In the round, the combination of core metrics, supplementary metrics and the written submission usually provided a sufficient evidence base for the panel members to come to a judgement in a clear and informed way, bringing to bear their professional expertise.

371. Non-reportable metrics were a significant problem in some cases and greatly reduced the panel's confidence in the robustness of decisions. It was a significant improvement to the process to be able to use the 'Not enough data' clause at the Model A rating meeting. This was used in five of the 38 cases. Generally, providers did not engage well with non-reportable metrics in their written submissions.

372. The panel was uncomfortable when lack of data led to defaulting to Silver and in many cases moved these subject areas to Bronze when assessing under Model B, feeling much more comfortable with the Bronze rating and the match to the rating descriptor for Bronze. Commonly, small number cohorts were a feature of further education colleges, and written submissions were relatively poorly argued, with weaker evidence. There was some evidence that colleges did not have a sufficiently clear understanding of the TEF assessment process and of the metrics.

373. The panel would recommend that stronger and clearer guidance is given to providers about the nature of evidence and impact. Panel members were not generally impressed by individual student quotes, when there was no supporting evidence to suggest how representative they were. There was also some scepticism about quotes from external examiners. Some panel members wished to give credit for sound mitigating actions to address weaknesses, even when impact may not be evident yet. This was particularly the case where providers were tackling complex student issues, for example relating to widening participation groups.

374. The panel would have liked to have seen stronger and more detailed coverage of PSRB accreditation issues where these applied. It would also be helpful if written submissions began with a clear statement of which courses are included within the subject area submission.

375. Employment data was not as reliable as it should be for two reasons. First, the rating did not distinguish between people already in employment and doing part-time courses and those who left higher education and then started employment. Second, lack of data on international students in all metrics except for the NSS, and exclusion of UK students working overseas, undermined the value of the metrics to the employer representative in particular.

Quality and robustness of the assessment process

376. Considering the three-step assessment process, the panel felt that the initial hypothesis at 1a set in train a process in which the burden of proof to shift from the initial hypothesis sometimes felt too heavy, and this was driving some weaker submissions to Silver when the panel felt they more appropriately matched the Bronze rating descriptor. When assessing under Model B, the panel was comfortable in taking an assertive view in the face of this issue, and so more confidently moved weaker submissions to a Bronze rating.

377. This issue of the 'anchoring effect' of the initial hypothesis may have also made it harder to move submissions to Bronze. On balance, the panel felt it had been more successful in 'spreading Bronze awards' than 'spreading Gold awards', especially under Model B. Under Model A, there was very little movement downwards from initial hypothesis to final rating.

378. The panel was concerned at the seeming predominance of Silver ratings under both models and would have liked to have awarded more Golds in particular. The panel referred to the ratings descriptors but felt that at subject-level, they set the bar too high for Gold and there were subject areas that we felt could potentially have been Gold but fell foul of the ratings descriptors.

379. The panel was comfortable with the range of expertise among its members. Some discussion occurred about the value of Office for Standards in Education, Children's Services and Skills (Ofsted) ratings in the education subject area. Several panel members had experience of overseeing education departments and felt clear about the significance of Ofsted ratings. However, some additional PSRB expertise would have been helpful.

380. For Model B, making 109 ratings decisions over effectively one and a half days put a great deal of pressure on the agenda and sometimes restricted time for discussion. The panel delivered on time, but did not think that this pace of decision-making would be appropriate for subject-level TEF proper.

Model A and B design considerations

Model A

381. The Model A approach to five-page submissions per subject area was preferable. Ideally, it would be helpful to have some standard institutional context material at the start of each subject area written submission. Where providers did include some introductory contextual material about the whole institution, this was well received by assessors.

Model B

382. Under Model B, if ratings could be awarded at the subject group level, rather than the subject area level, this would improve the efficiency of assessment and could also help with the challenges of scaling up. However, it would be less likely to be helpful in informing students of the level of teaching excellence.

383. Under both models, it would be helpful to provide clear information to assessors about which courses are covered in the subject area.

384. The panel would favour a mandatory requirement that PSRB accreditation is explicitly covered in the subject area written submission.

Subject-specific considerations

385. There was not a great deal of subject-specific discussion during the panel's deliberations. There was very little discussion of QAA subject benchmark statements, for example.

386. In the written submissions there was very little presentation of any distinctive ethos around pedagogy. Some panel members felt that the emphasis on the standard TEF Year Two metrics risked a sterilising effect upon higher education pedagogical approaches. Subject area written submissions rarely discussed their 'signature pedagogies.'

387. The areas which often prompted most questioning were education and architecture. The concerns with education included the level of importance placed on Ofsted ratings and the significance of employability data for part-time students. For architecture, continuation and NSS metrics seemed to be affected by the structure of the architecture degree. It was helpful to have architecture specialists on the panel who were able to interpret these issues for the panel.

Segmented panel member views

Students

388. Student panel members played a full part in the deliberations and there was no discernible distinction between the contributions made between student and academic panel members. Student members felt that employability data was overemphasised in the assessment process and would have liked to have seen more emphasis on WP data. Student members suggested that it would be helpful to have an explicit step in the assessment process to consider WP and student voice issues.

Employers and PSRBs

389. The employer panel member played a full role in the discussions. The employer representative found the lack of data on international students concerning. International students are only included in the NSS data. They are excluded from DLHE data and are not in the LEO data. The time and effort that international employers invest in careers fairs and links to higher education providers could be undermined if providers are not judged on metrics on students from overseas or placement of UK students in jobs overseas.

390. The panel had no PSRB representative. We were able to cover PSRB issues, and there was enough expertise across the panel members, but a PSRB representative might have strengthened the treatment of these issues.

Feedback on submissions

391. Providers should be given more explicit guidance on the following:

- addressing evidence gaps where there are non-reportable metrics
- evidencing impact of performance improvement initiatives
- actions in response to actively engaging with students and their representatives

- the significance of PSRB accreditations.

392. It would be helpful if providers were supplied with more detailed guidance that they should explicitly address information gaps left by non-reportable metrics, and also where they consider that the standard TEF approach to metrics is not appropriate for their particular context. Under Model A, it would be helpful if providers were required to explicitly address the question of the exceptionalism of the subject area in the written submission, if an exceptionalism-based model is to be pursued.

393. Although providers commonly structured their written submissions using the three main sets of criteria – teaching quality, learning environment, and student outcomes and learning gain – there were many written submissions that did not deal explicitly with all the 10 criteria that fall within these headings. Panel members felt it would be useful in the assessment process to have the 10 criteria easily to hand to structure assessments in a more detailed way. It was noted that there was very little narrative on the use of technology to support learning and teaching.

394. Further education college submissions tended to be poorer in quality than those submitted by higher education institutions.

395. Student engagement was generally not thoroughly dealt with in submissions, with only a few institutions evidencing good practice. The students across subject panels met together separately and will be reporting separately on how consideration of student voice might be better and more robustly incorporated within the TEF assessment process.

396. Some providers struggled with the ambiguous use of the term ‘student engagement.’ Against the TQ1 (student engagement) criterion, they discussed student representation and student voice rather than student academic engagement in their studies.

397. Some panel members were concerned about the overall judgements when submissions had distinctive patterns of performance that varied significantly between full-time and part-time students. They would have preferred to be able to give different ratings for full-time and part-time provision.

Potential impacts

398. There was a concern among the student panel members, which was widely shared, that the focus on outcomes split by WP characteristics could potentially provide a disincentive for providers to recruit from underrepresented groups. Evidence of satisfaction, attainment and outcomes among WP groups therefore need to be considered alongside the contextual information on student intake, and any strategies to strengthen representation of underrepresented groups in the providers’ student bodies.

399. The disadvantage suffered by further education colleges set out in paragraph 373 is also problematic because, according to the metrics, they appear to be relatively strong on widening participation.

400. The relative lack of emphasis on student voice, in TEF guidance and in written submissions, is a missed opportunity to use the TEF to strengthen the incentives for institutions

to engage with students in partnership in the management and enhancement of their education.

Additional observations

401. The panel operated well and there was a good range of expertise and highly informed discussion to arrive at ratings. The sessions covering feedback on the assessment process were particularly productive, and there was a high degree of consensus around the majority of the insights generated from the exercise.

402. It would be useful to annually review the guidance for TEF assessors to improve guidance on assessment and rating and promote consistency across panels.

403. The arrangements and administrative support for the panel's work were very good. The OfS TEF Team provided helpful guidance. The TEF officer provided very helpful administrative support to the chair and deputy chair to assist with the allocation of assessments.

Report author: Prof Neil Ward, Chair of the Social Sciences Panel

Table 7: Social Sciences Panel members

Chair	
Prof Neil Ward	Deputy Vice-Chancellor and Pro Vice-Chancellor (Academic Affairs), University of East Anglia
Deputy Chair	
Diarmuid Cowan	Former Students' Union President, Economics School and Class Representative, Heriot-Watt University
Panel members	
Emma Beenham	Students' Union Academic Affairs Officer, Aberystwyth University
Prof Alvin Birdi	University Academic Director of Undergraduate Studies, University of Bristol
Prof Joanna Bullard	Associate Dean (Teaching) and Professor of Geography, Loughborough University
Prof Debby Cotton	Head of Educational Development and Professor of Higher Education, University of Plymouth
Prof Joelle Fanghanel	Pro Vice-Chancellor, University of West London
Prof Dilly Fung	Pro Director (Education) (from May 2018), London School of Economics and Political Science

Catherine Higgs	Associate Head of Faculty (Construction), University College of Estate Management
Cath Holmstrom	Deputy Head of School and Head of Department, University of Brighton
Prof Christina Hughes	Provost, Sheffield Hallam University
Hien Le	Former Students' Engagement Assistant, University of Bath
Kate Mori	Head of Teaching and Learning, Hartpury College
Melissa Owusu	Former Students' Union Education Officer, University of Leeds
Dr Andrew Roberts	Dean of Education and Students, Cardiff University
Prof Ann Shelton Mayes	Executive Dean (Student Experience), University of Northampton
Marie Staunton	Chair of Crown Agents, Crown Agents
Jonathan Stephen	Students' Union Education Officer, University of Huddersfield
Jan Thompson	Faculty Association Representative, the Open University
Prof Malcolm Todd	Pro Vice-Chancellor (Academic and Student Experience), University of Derby

QAA TEF officer: Irene Ainsworth

Piloting TEF at a subject level: Widening participation report

Methodology

404. The Widening Participation (WP) experts were full main panel members and advisors to the panel. Both experts had a caseload of provider-level assessments for pilot Models A and B (12 institutions each, with some overlaps for comparative purposes). In addition, they observed a limited number of subject panels in both Model A and B, and contributed to the main panel discussions particularly in relation to widening participation and student diversity issues. The experts also provided specialist advice to the main panel and, to a lesser extent, to some of the subject panels.

405. The experts used WP data tables supplied by the OfS team[9] and sector knowledge to select the institutions they would assess under Model A and B. There was regular contact and feedback between the TEF officers and OfS team and the WP experts. Throughout the process notes were taken by both experts on widening participation issues.

Observations

406. During the subject panel meetings, different meeting styles were observed. This was not considered to be detrimental to the final assessment outcome; however, it was noted that additional time and debate did allow for all considerations to be taken into account by the panels. For example, some panels were more swayed by the metrics, while others engaged in a more holistic judgment, allowing for greater consideration of WP issues. At times an overreliance on the metrics was observed by the WP experts with a reluctance to use judgment.

407. It was observed that there was a mixed level of understanding by both the main and subject panels on what was contained within the benchmarked data. The experts were not convinced that all panel members fully understood the significance of the benchmarked data and what is omitted. For example, sometimes it was assumed that the benchmarks incorporate all aspects of diversity and performance, and so any deviation from the benchmarks should be viewed negatively. This approach does not sufficiently take account of a very diverse student population, incorporating aspects of diversity (such as commuting and indices of multiple deprivation) which are not benchmarked, or the cumulative impact of large numbers and intersectionality (e.g. intersections between POLAR and age). In some instances suggestions to consider the diversity of the student population and what the institution is doing to address the outcomes of students with these characteristics was resisted by saying this is all taken into account in the benchmarks. Similarly there was a lack of understanding that there is no geographic element to the employment benchmarks.

408. The experts found the maps to be helpful in some of the assessments; however, feedback from other panel members indicated that they were not used widely in the assessment process.

[9] OfS analysts produced a table containing existing information reformatted to be most useful to the WP experts – it presented contextual data alongside summary data on split metrics for each provider.

409. It was difficult for the subject panels to fully understand the WP context from just the subject submission, especially as small numbers often resulted in the suppression of split metrics data in relation to diversity. This is particularly the case in Model B, and the WP experts are concerned that much of the key WP-related data is lost at subject-level, particularly when the numbers are low enough to be unreportable. It is also noted that lower Z-scores (indicating low levels of statistical significance) in the small subject areas mean that there is little confidence in the metrics being presented.

410. A number of provider submissions were weaker. Some submissions did not draw any or sufficient attention to the composition of their student population (either assuming it was irrelevant or conversely, that it would be known). Submissions regularly omitted to address the concerns within the metrics, overall or in relation to split metrics, diversity or outcomes, or failing to provide evidence to support the claims. It is known that the sample of the institutions was limited; however, a number of these institutions were either alternative providers or further education providers.

411. We noted that the level of understanding and expertise varied between panel members and between panels and groups; furthermore there was not always a common or shared understanding of the issues. This could relate to submissions being treated differently and inconsistency in the judgements. This was addressed well in TEF Year Two given the extended time for discussion and debate by the main panel, and so steps to preserve 'WP consistency' should be taken in all future TEF assessments.

Recommendations

412. It was not considered necessary or practical for each subject panel to have a WP expert (this would create further challenges regarding consistency). The WP experts believe it is important that all members of the subject panels develop a proficient understanding of the WP issues within the sector and know how to address them within the assessment process. It is proposed that a liaison model is adopted, where a member of each subject panel (which could be the chair or deputy) is a link person with a small group of two or three WP experts on the Main Panel.

413. The link person's role would be to identify and draw attention to potential WP issues, and the WP experts would be able work with these liaison members to answer queries and share best practice across the subject panels.

414. The link people across subject panels would require some training. This could be supplemented by a WP-related step in the model, which could relate to access and participation plans.

415. A WP discussion board should be established on the QAA system to capture WP issues and share best practice.

416. The liaison model could be replicated for the employment experts.

417. The WP experts should be able to attend any subject panels and provide input where necessary. It was not helpful to the panels for WP experts to have observer status. Subject panels should also be able to refer a provider submission and metrics to the WP experts for consideration.

418. Further guidance on constructing the provider submission may support some providers to better demonstrate the makeup of their student population, and what actions they are undertaking to address their WP issues. A basic template has the potential to ensure the core areas are outlined within the submission and may help the provider to highlight areas of best practice.

419. The WP experts could also provide training to providers, either online or through a conference format.

420. Where a provider is either single-subject or has a small number of students, the guidance could outline a ‘less but equal’ methodology for both the provider and the assessor.

421. It is recommended that further training is given to all panel members and assessors on a number of key areas. This includes guidance on what each benchmark contains and the importance of taking intersectionality into consideration.

WP experts and report authors: Prof Liz Thomas and Ross Renton

Employment experts' report

Role of the employment experts

422. We would argue in favour of a continued role for the employment experts in future TEFs. On the main panel, experts fulfilled a clear role, providing context and information relating to employment issues in certain cases, such as labour market observations for particular subjects or in particular regions. This proved useful especially for smaller providers whose provision was dominated by particular subjects or whose students worked in particular regions following graduation. As for the subject panels, employment experts could perform a similar role on such panels, though this would require an increase in numbers to cover all panel meetings. If their role on subject panels is to be restricted to sampling different panels on certain days, as was done in the current TEF, then we recommend that this should done on the first day of panel meetings wherever possible, when more cases are discussed.

423. An alternative to attending subject panel meetings (either all meetings or a sample) would be for the employment experts to be available to be contacted by subject panels when particular issues related to employment emerged in panel discussions, and to offer information to help the panel members with their deliberations. We feel this would be the most effective use of the experts' time, while providing the most assistance to the subject panels.

424. We recommend that employment experts should continue to receive their own caseload of cases to consider, so that they are aware of the issues that can arise and the decision-making processes undertaken by panel members. Ideally the cases considered should be ones that raise interesting issues regarding employment. These can be either self-selected, recommended by the OfS, or a mixture of the two.

Use of employment measures in TEF classifications

425. Based on our experience of the first subject pilot, we considered more broadly how employment metrics might be developed for subject-level TEF.

426. The first issue is whether employment measures should be used at all, when the aim is to measure the quality of teaching that students receive. The argument against is that employment outcomes are a very indirect indicator of teaching quality, determined as they are by a range of other factors. If the outcome measure is beyond the direct control of the agent being evaluated, this reduces incentives to alter performance to try to affect outcomes. The arguments in favour of using such measures are that employment outcomes are observable and quantifiable, and reflect at least one of the aims of higher education. Alternative metrics, such as measuring students' skills in some way, observing and rating actual teaching, or relying only on subjective ratings such as the NSS, all have obvious limitations in terms of feasibility, expense or desirability. We therefore agree that there should be a continued role for employment statistics in the future TEFs.

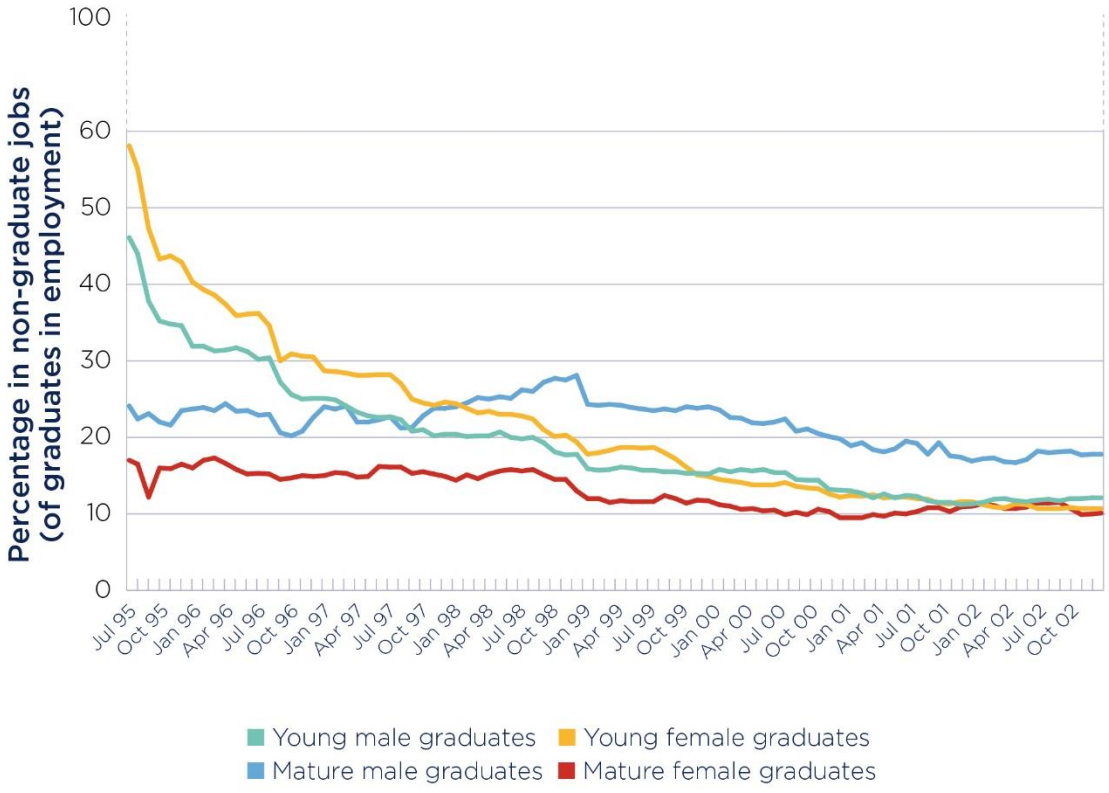
427. However, one question we raise is whether having two of the six metrics relating to employment may be too many, particularly given that they both carry a full weight. An institution that does well on employment issues, and attracts two positive flags, is then well on the way to achieving a Gold-rated initial hypothesis. Maybe the 'any employment' measure could be dropped, or reduced in weight at least, since it is 'highly skilled employment' that is the desired

outcome for students and policy-makers. In a labour market that evidence suggests is becoming more polarised into high-level and low-level jobs over time, with declining numbers of intermediate-level jobs, there is a risk that the value of the non-graduate job alternative is likely falling on average as more graduates who fail to obtain graduate level employment find themselves in lower-level jobs, meaning just being in 'any' job is worth less on average now.

428. In addition, refinement of the 'highly skilled employment' definition would be useful, since the Standard Occupational Classification 1 to 3 definition is rather blunt for these purposes. Alternative classifications of 'graduate jobs' are available. We recommend that consideration should be given to these.

429. Moving on to sources of employment data, the focus on the six-month DLHE in the current TEF should be moved to the new 15-month DLHE as soon as the latter is available. Measuring employment outcomes for graduates six months after graduation is too early to pick up their settled destination in many cases. Figure 1 charts the movement of graduates from the class of 1995 over a subsequent seven-year period. While this is a historical picture, we have no reason to believe that the situation for leavers from higher education is vastly different nowadays.

Figure D1: Movement out of non-graduate jobs (percentage of all in employment) by age at graduation and gender (1995 graduates from 38 universities, 1995 to 2002)[10]



[10] Purcell, K., N. Wilton and P. Elias (2003) 'Older and Wiser? Age and experience in the graduate labour market' Researching Graduate Careers Seven Years on, RP2 University of Warwick, Institute for Employment Research (https://warwick.ac.uk/fac/soc/ier/research/completed/7yrs2/rp2.pdf).

430. There should also be continued engagement with the LEO data, particularly as that source will continue to develop and improve. At the moment, the main advantages of LEO are the ability to track individuals over time, rather than focusing only on their initial entry to the labour market, and the detailed and accurate earnings information. For now, these advantages need to be counterbalanced against the disadvantages, which include the time lag that means older cohorts of graduates are being considered, the absence of data on occupation, and the lack of other background characteristics (particularly region and family background) with which to benchmark the labour market outcomes. Avenues should continue to be explored for overcoming these disadvantages, for example the possibility of coding occupation from tax returns, or of matching in data on variables such as region and occupation from alternative existing or forthcoming data sources, for example the 2011 and the 2021 Censuses of Population, or the Annual Survey of Hours and Earnings.

431. Finally, whatever the source of employment data, it is crucial that it is benchmarked as well as possible. Looking at the outcomes in the current round of TEF, positive flags indicating performance above benchmark are consistently achieved for the 'highly skilled employment' measure by Russell Group and similar research-intensive so-called 'elite' universities. Despite the higher benchmarks set for such institutions, it therefore seems that the benchmarks are not doing a good job of controlling for the types of students who typically attend these universities. In particular, the local area-based measures of socio-economic status and family background that are currently used (POLAR and indices of multiple deprivation) are not sufficiently fine-grained to pick up differences across families. Family-specific measures of socio-economic status, mapped to employment data, should therefore be sourced if at all possible.

Employment experts and report authors: Prof Peter Elias and Prof Steven McIntosh