

Office for
Students



Teaching Excellence and Student Outcomes Framework

Findings from the first subject pilot,
2017-18

Reference OfS 2018.43

Enquiries to TEF@officeforstudents.org.uk

Publication date 22 October 2018



Contents

Introduction and key points	3
How the pilot was conducted	6
Main Panel Chair's report	10
Synthesis of findings	15
Summary ratings and analysis	42

Available alongside this document on the OFS website:

Annex A: Feedback from student representatives

Annex B: Panel reports

Annex C: Evidence from providers

Annex D: Cost survey

Annex E: Teaching intensity

List of abbreviations



Introduction and key points

Purpose

1. This report summarises the results and key findings from the 2017-18 Teaching Excellence and Student Outcomes Framework (TEF) subject pilot. The pilot was delivered initially by the Higher Education Funding Council for England (HEFCE) and then, from April 2018, the Office for Students (OfS), according to the Department for Education's (DfE's) specification. For simplicity, we will refer to the OfS in place of HEFCE throughout the report.
2. An earlier version of this report was submitted to DfE in July 2018 to inform development of the government response to the consultation on subject-level TEF. It has also subsequently informed the development of the guidance for the second subject pilot (2018-19). These two documents set out our response to the findings in this report, including the changes that will be tested in the second year of pilot as a result of the findings.

Background

3. The TEF was introduced by the DfE in 2016 to recognise and reward excellent teaching in UK higher education.
4. The Office for Students was established by the Higher Education and Research Act 2017 (HERA)¹ and became fully operational in April 2018. The TEF is the scheme that the OfS has adopted under HERA section 25(1) as a sector-level intervention to promote excellence in teaching and outcomes.
5. The current TEF assesses 'provider-level' undergraduate provision. It aims to shine a spotlight on teaching and give potential applicants clear, easy-to-understand information about where excellent teaching and student outcomes can be found. The May 2016 White Paper 'Success as a knowledge economy' set out the intention to undertake TEF assessments at a disciplinary (subject) level, thus making TEF even more useful to potential applicants¹.
6. Through two years of piloting we aim to ensure that the future design of subject-level TEF appropriately encompasses the full variety of higher education provision across the UK. Colleges and universities based in the devolved administrations were represented in the sample of providers and on the panels.
7. The key concern of the first pilot was to test how two potential models generated ratings at both subject and provider-level. We sought to:
 - robustly compare the feasibility of the two models
 - assess whether the full range of detailed design and delivery aspects of the pilot were fit for purpose
 - consider implications for delivery at scale.

See <https://www.gov.uk/government/publications/higher-education-success-as-a-knowledge-economy-white-paper>.

8. In addition, the feasibility of a standardised measure of teaching intensity was also tested².
9. Implementation followed both the DfE's TEF subject-level pilot specification, published in July 2017³ and the 2017-18 'Year Three provider-level' specification⁴ which was subsequently released in October 2017. This report assumes readers are familiar with both specifications and how the TEF has developed to date⁵.
10. Alongside the first subject pilot, the DfE conducted a technical consultation with the sector⁶ and commissioned student research. The student research tested the utility of subject classification systems with applicants, and explored applicant and student views on the relative importance of teaching quality and student outcome factors in applicant decision-making and the quality of student experience⁷.
11. The consultation, student research and findings from the first subject pilot (as set out in this report) have collectively informed the government response and OfS guidance for the second subject pilot.
12. The first pilot was a developmental exercise and the findings in this report relate to the overall models and methods of assessment, not to individual providers that took part. Confidentiality was a necessary requirement in order for all pilot participants to feel comfortable about fully engaging with the process. Participants received their own indicative ratings (but not those of other participants) and agreed they would not publish their own ratings. Therefore none of the indicative ratings generated are published in a way that identifies individual providers, nor will they impact on applicant choice, fee caps, or any other OfS activities.
13. The OfS team would like to thank the DfE TEF team, the 50 higher education providers, the 141 panel members and the Quality Assurance Agency for Higher Education (QAA) staff for their dedication, professionalism and constructive support throughout. Particular thanks go to the Main Panel Chair, Professor Janice Kay, for her leadership, insight and advice.

² Guidance on teaching intensity can be found at <https://www.officeforstudents.org.uk/advice-and-guidance/teaching/piloting-tef-at-a-subject-level/further-technical-information/>.

³ See <https://www.gov.uk/government/publications/teaching-excellence-framework-subject-level-pilot-specification>.

⁴ See <https://www.gov.uk/government/publications/teaching-excellence-and-student-outcomes-framework-specification>.

⁵ See www.officeforstudents.org.uk/advice-and-guidance/teaching/piloting-tef-at-a-subject-level/.

⁶ See <https://consult.education.gov.uk/higher-education-reform/teaching-excellence-and-student-outcomes-framework/>.

⁷ See <https://www.gov.uk/government/publications/teaching-excellence-framework-and-informing-student-choice>.

Key points

14. This report outlines the following key findings:

- a. Ratings were successfully generated at provider and subject-level in both models using the existing provider-level framework.
- b. While the evidence, criteria and descriptors used in current provider-level TEF can be applied in a subject-level exercise, a number of refinements could be made to enhance their utility to panel members and meaningfulness for students.
- c. Both models feature design elements that were intended to reduce burden but ultimately added to the complexity of the exercise and did not produce robust ratings for all subjects in their current form. Panel members and providers expressed support for a more comprehensive model, drawing on the best elements of Model A and Model B.
- d. The method of assessment based on combining metrics and submissions worked well but there were limitations in using the data at subject-level. There was a tendency for metrics to 'default' to a Silver initial hypothesis and panels found they were unable to reach judgements, or were not confident in the judgements they made, where there was a sparsity of data arising from small subject cohort sizes and missing data sources.
- e. The profile of provision at subject-level can shift rapidly, and both providers and panels identified cases that were inappropriate to assess. Clear criteria and processes will need to be developed for identifying provision that should be out of scope for subject-level assessment.
- f. Student input in submissions at subject-level was valuable where present but was patchy, and efforts will be needed to enhance this in future.
- g. For providers, there was little difference in the costs of the two models.
- h. Assessment by panels each covering a group of subjects worked well, but there will be major implications of scaling up to a full scale exercise.
- i. Grouping subjects at the second level of the Common Aggregation Hierarchy (CAH2) level worked well, providing the right level of aggregation. However some refinements to the classification could be made to better reflect the diversity of provision in UK higher education.
- j. The issue of adequately representing interdisciplinary subjects which map to two or more CAH2 categories will require further consideration. Improved contextual data will enable such provision to be better understood and more can be done to enhance panel expertise of interdisciplinary provision.
- k. The use of accreditation data proved useful, however there is a need for improved contextual data to enable this is better interpreted.
- l. The teaching intensity measure was a significant burden to providers and was not found useful by panel members.

How the pilot was conducted

The models we tested

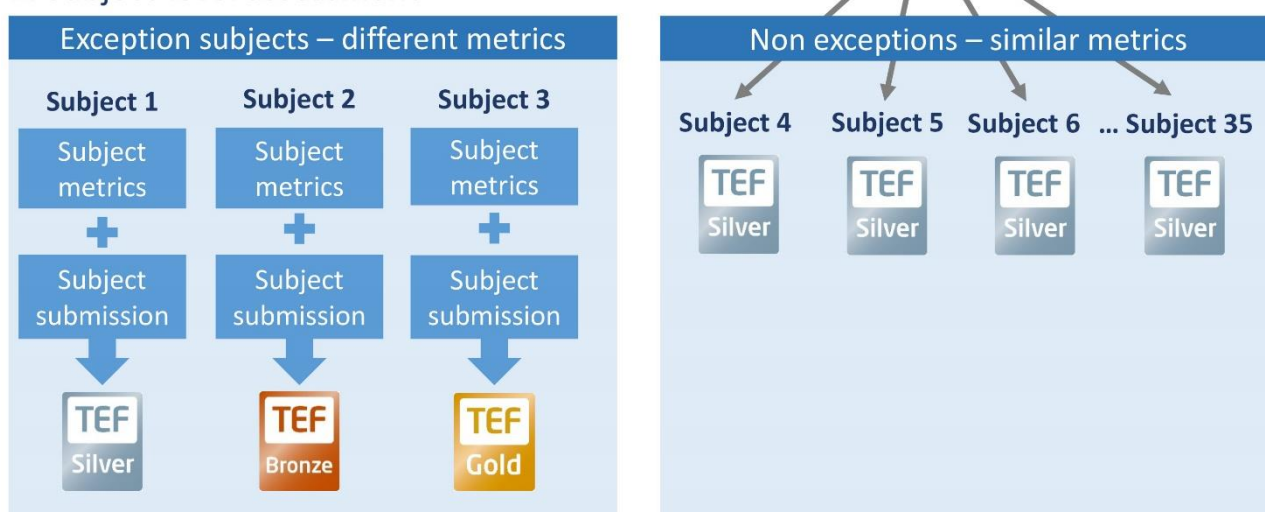
15. Two models were tested in the first subject pilot. For both, assessment was based on the same criteria and evidence used in TEF currently. Importantly, TEF ratings were produced for each subject, and for the college or university as a whole.
16. Model A is a 'by exception' model. Subjects were identified as 'exceptions' where the metrics perform differently to the provider-level metrics. These exception subjects were fully assessed and received a rating of Bronze, Silver or Gold. Non-exception subjects were not assessed, and received the same rating as the final provider rating as shown in Figure 1.

Figure 1: Model A assessment⁸

1. Provider-level assessment



2. Subject-level assessment

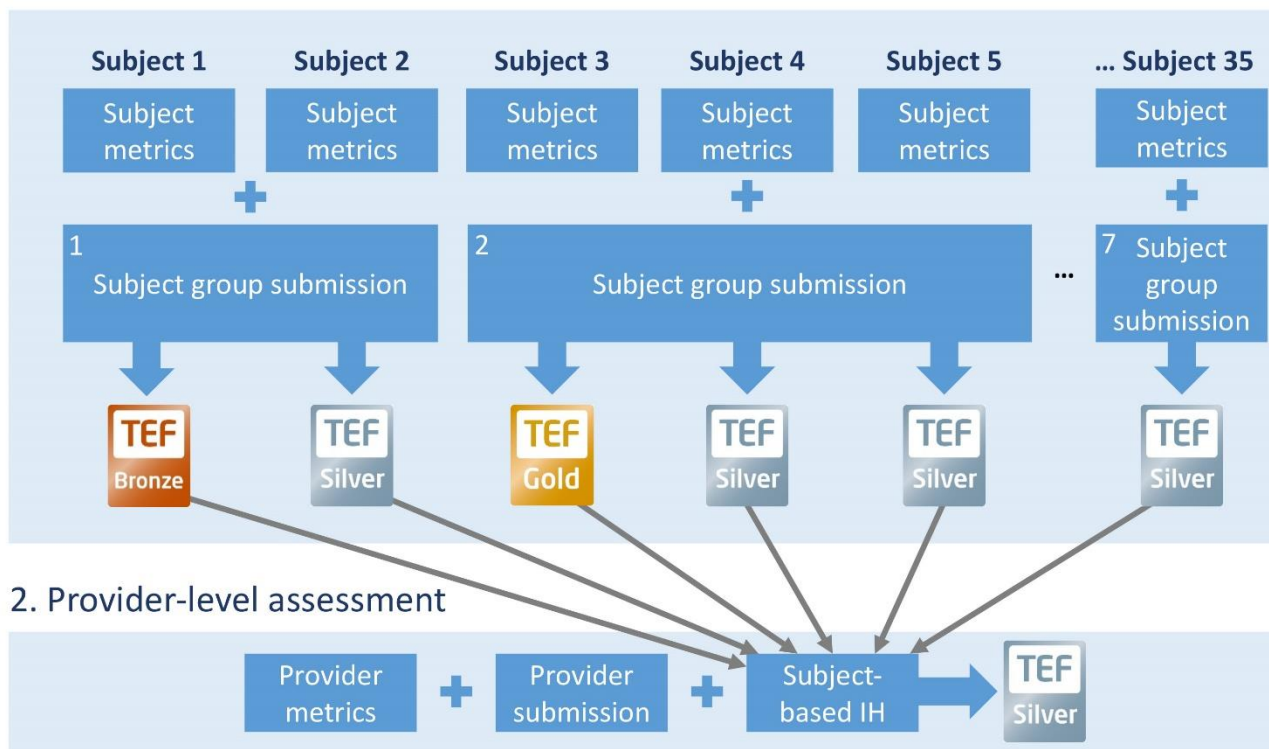


17. Model B is a 'bottom-up' model. All subjects were assessed as part of a 'subject group' submission but with separate metrics for each subject, and each subject received a TEF rating of Bronze, Silver or Gold. The subject ratings then fed into the provider-level assessment through a subject-based initial hypothesis, as shown in Figure 2. Additionally In Model B, it should be noted that provider-level submissions had shorter page lengths and focussed on three of the 10 TEF criteria (Teaching quality criterion 2 (TQ2): Valuing teaching, Learning environment criterion 1 (LE1): Resources and Student outcomes and learning gain criterion 3 (SO3): Positive outcomes for all).

⁸ Figures 1 and 2 were developed from diagrams originally produced by the DfE for use at briefing events.

Figure 2: Model B assessment

1. Subject-level assessment



Who took part in the pilot?

18. A total of 50 colleges, universities and other types of higher education provider took part in the first subject pilot. They were selected to reflect the diversity of providers across the UK. The full list of participants and the models they participated in can be found in Table 1.

Table 1: participants in the first year of TEF subject pilots

Model A participants	Model B participants	Participants in both models
Bath Spa University Chichester College Futureworks Glyndŵr University Hertford Regional College Imperial College London Kingston College Liverpool John Moores University London Metropolitan University	Abingdon and Witney College Arts University Bournemouth Furness College Glasgow School of Art Kaplan Open Learning (Essex) Limited Leeds Beckett University Leeds Trinity University London School of Economics and Political Science	Bishop Grosseteste University Bromley College of Further and Higher Education De Montfort University Nelson College London Limited Pearson College Limited St George's, University of London University of Birmingham University of Bradford University of East Anglia The University of Law

Model A participants	Model B participants	Participants in both models
London School of Management Education Limited Open University Ravensbourne University of Bedfordshire University of Cambridge University of Essex University of Hull University of Nottingham University of St Andrews Warwickshire College Group	London South Bank University New College Durham School of Oriental and African Studies South and City College Birmingham Swansea University Tyne Coast College University of Exeter University of Hertfordshire University of Leicester University of Portsmouth University of Westminster	University of Northumbria at Newcastle University of Wolverhampton

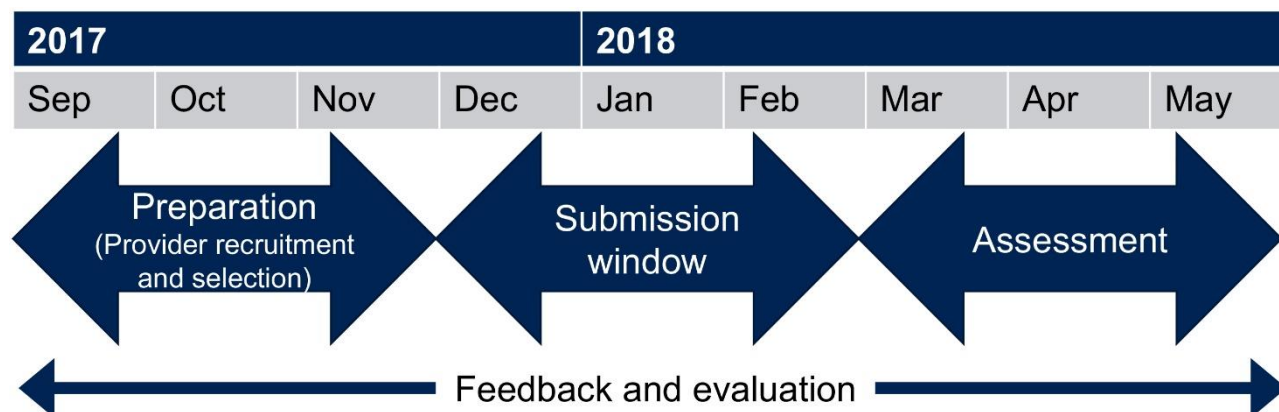
The panels

19. Over 140 panel members carried out the pilot assessment process. They were selected for their standing in the higher education sector, expertise, and commitment to excellence in teaching.
20. The Main Panel decided ratings at provider-level. The seven subject panels decided ratings at subject-level. Each panel consisted of student representatives, academics, and representatives from employers, widening participation experts and professional, statutory and regulatory bodies (PSRBs).
21. The Main Panel was chaired by Professor Janice Kay, Provost and Senior Deputy Vice-Chancellor, University of Exeter with Professor Helen Higson, Provost and Deputy Vice-Chancellor, Aston University as deputy-chair. Each subject panel had an academic chair and a student deputy chair who were also members of the Main Panel.

Timings

22. The pilot ran alongside TEF Year Three but did not interact with the provider-level exercise in any way. As shown in Figure 3, the pilot consisted of three broad phases: preparation, submissions, and assessment. Providers participating in the pilot received subject-level metrics at the beginning of December 2017 and had a three-month window to complete their submissions. Assessment took place between March and May 2018. Model A and Model B assessment processes were carried out separately but conducted by the same panel members to inform a comparison of the models.
23. The exercise featured 17 briefing and training events, attended by over 500 individuals, and 20 assessment and feedback events that took place over a total of 22 days.

Figure 3: 2017-18 pilot assessment timeline



Evaluation

24. Ongoing evaluation and refinement will be required to ensure that, overall, the move to subject-level TEF delivers the intended policy outcomes (to better inform students' choices about what and where to study, to raise esteem for teaching, to recognise and reward excellent teaching, and to better meet the needs of employers, business, industry and the professions). In the first pilot, the OfS focused on whether the Model A and B processes and their implementation had the potential to effectively meet these policy goals and what would likely be the positive and negative impacts for higher education providers, provision and students. An iterative and formative approach to the pilot was taken and emerging findings were discussed by the DfE and the OfS throughout. This publication includes a report from the Main Panel Chair, a synthesis of our final findings from across the exercise, and an analysis of the indicative ratings generated. Additional evidence from the pilot's evaluation strands are included in this document's annexes:

- a. Feedback discussion was built into the agenda for panel assessment meetings. Subject panel feedback was collated in subject panel reports, authored by the panel chairs. A list of topics for panels to investigate was provided at the start of the assessment process to ensure consistency across panels. A Main Panel report was produced by the Main Panel Chair, Professor Janice Kay. Her report (pages 11-15) provides a thematic analysis of the subject panel reports and feedback from panel observers. All subject panel, widening participation expert, and employment expert reports are provided at Annex B.
- b. Initial feedback from participating providers was collected at pre-assessment briefing events by OfS officers. Final feedback was captured in a post-submission survey and post-ratings in-depth interviews. An analysis of provider feedback is given at Annex C. Participating providers were also asked to complete a cost survey to facilitate a comparison of the cost of each model. An analysis of the cost survey results is given at Annex D.
- c. In May 2018, the OfS carried out two focus groups with student representatives, and a further focus group with student pilot panel members. An overview of feedback from these focus groups and a report produced by subject panel deputy chairs are given at Annex A.
- d. The development of the teaching intensity measures and analysis of the data captured are given at Annex E.

Main Panel Chair's report

The Main Panel Chair's report, authored by Professor Janice Kay, synthesises key findings and recommendations from the Main Panel, subject panels, and widening participation and employment expert panel members. The seven subject panel reports, widening participation expert report and employment expert report are all included at Annex B.

Methodology

25. Throughout the pilot we captured panel members' feedback on:

- how the design and specification of each model influenced the robustness of the assessment outcomes
- their wider experience of the assessment process, the utility of the information available and what other information would have been beneficial
- the key challenges and opportunities they perceived.

26. The approach was intentionally formative and we used panel members' training and calibration meetings to refine and clarify procedures. During the assessment phase OfS and QAA staff and other panel members acted as observers of panel process effectiveness and consistency. This led to iterative consideration and feedback in the panels on issues arising and potential solutions.

27. The resulting feedback was captured and summarised in the seven subject panel reports, which each drew on a topic list developed by the TEF team to support subject panel chairs in compiling their panels' feedback on each model following the subject panel meetings. The same topic list was used by the TEF team to capture feedback from Main Panel meetings and to assist the Main Panel Chair in compiling her high-level report on the pilots, incorporating the Main Panel's reflection on the models and recommendations for the second year of the pilot.

28. Student members contributed to feedback and evaluation throughout the process, and additionally a focus group was held with a subset of student panel members at the end of the subject panel meetings to discuss student-specific issues, summarised in a report drafted by the student subject panel deputy chairs. The widening participation and employment experts on the Main Panel also had a role in providing specific feedback on how their area of expertise was handled in the subject pilot, noting that TEF panels benefit from having members with an in-depth knowledge of these areas. Both sets of experts produced short reports summarising the key issues identified and recommendations for the next year of pilots.

Main Panel Chair's report

29. The second year of the subject pilots should be taken as an opportunity to test refinements to a single refined model which draws upon the best elements of both models. The model should be tested with a scaled-up cohort with particular attention paid to the scalability of the exercise ahead of roll out of subject-level TEF from autumn 2019 – this should include consideration of the timings of the TEF reapplication cycle.

Robustness of ratings

30. Members across the panels felt that robust ratings were generated, with an important caveat at subject-level: where metrics data was sparse, for whatever reason, panels had less evidence to inform their judgement and felt less confident in awarding a rating to these subjects. The opportunity to carefully agree 'No rating' for a minority of subjects, where panels felt evidence was missing across both metrics and submission, increased confidence in the process and the robustness of ratings awarded.

Model preference and design considerations

31. All Main Panel members expressed a preference for Model B, although neither model was felt to be appropriate for roll-out in full subject-level TEF in its current form.

32. The fullness of assessment in Model B made it a more comprehensive and reliable source of information for applicants. While the subject group submission and the subject-based initial hypothesis were weaknesses in Model B's design, overall panel members saw more opportunities for tweaks to improve this mode.

33. Model A was felt to be lacking in its provision of information to applicants and potentially misleading, with concerns around the unknown element of non-exception subjects that would inherit the provider-level rating with little interrogation. One of the key benefits of TEF to date, and of the future benefits of subject-level TEF, is increasing the focus on teaching enhancement. Panels felt that this would not occur across the board if only a sample of subjects were fully assessed. Non-exception subjects would not have an associated submission, making it difficult for students to make fully informed choices about a subject across institutions.

34. Subject panels' views aligned with the above, with the strengths of Model A at subject-level (focused subject submissions) theoretically able to be utilised in Model B, while the weaknesses of Model A (assessment only by exception) were much more difficult to reconcile.

Quality of the evidence

35. A significant concern at subject-level was the number of cases of non-reportable metrics, in evidence for a variety of reasons including small numbers, new courses, closed courses and teach-out of provision. A majority of subject panels were in favour of a minimum cohort size for subjects to be assessed and given ratings, with some form of 'provisional' award to be given to cases that did not meet the threshold in the real exercise. This was vital in terms of maintaining credibility in the methodology.

36. The measure of teaching intensity, piloted in five subjects, was not found to add value in panels' assessments of those subjects and was used very infrequently across the board.

The assessment process

37. The Gold, Silver and Bronze rating descriptors did not seem entirely appropriate for subject-level ratings and could be tweaked at both subject and provider-level. There was support within the Main Panel for keeping subject-level and provider-level submissions, but decoupling them, potentially extending to using different criteria to judge them. This arose from a majority of Main

Panel members rejecting the SBIH and the heavy influence it was felt to have on provider-level judgements.

38. Panels found that there tended to be 'stickiness' of the metrics in assessments, and a better balance between metrics and submissions is needed to promote the holistic judgement of providers and their subjects. This was particularly the case at subject-level in Model B, where a lower proportion of subjects were moved up than in Model A, and the panel felt that this was due to the methodology, rather than reflecting the genuine spread of excellence. Clearer targeted individual submissions are required, and more comprehensive training of panel members in completing a holistic assessment will help. More consideration is needed about how the initial hypothesis based on metrics should function and how panels could make more secure use of the subject written submission.
39. Generally, level 2 of the Common Aggregation Hierarchy (CAH2) subject classification worked for subject-level assessments, although most panels identified one subject in their subject group which was a 'mixed bag' of courses or did not align entirely naturally with the rest of the group.
40. Evidence of impact was the key indication of good quality submissions at both provider and subject-level. The best submissions addressed both strengths and weaknesses in their data, showed an understanding of their mission and their students, and were able to identify where interventions had resulted in a clear positive change. Some panels felt that specifics in subject discipline did not come through strongly enough in submissions, and ways to demonstrate this could be made more explicit in the criteria at subject-level.
41. Widening participation and employment experts on the Main Panel also made a number of recommendations for the process. The role of both widening participation and employment experts at subject-level should be a consultative one, where the experts are a resource to be drawn upon when subject panels have particular widening participation or employment related queries regarding a particular provider. Panels would benefit from further training on widening participation issues, particularly where split data becomes less reportable due to small numbers at subject-level.

Potential impacts

42. Further work is needed on how subject TEF will inform student choice – both current models could result in difficulties in the way subject ratings would be presented to potential applicants, and getting this right is critical to the integrity of TEF. The messaging around subjects that do not have sufficient data for assessment must be carefully considered. How subject-level information is communicated becomes vital and should be a consideration in the second subject pilot (2018-19).
43. There remained some concern among subject panels and the Main Panel that subject TEF could lead to unintended consequences and stifle innovation. As the chair of the Humanities Panel reports:

'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’.

44. Another consideration at subject-level was the lack of data for new courses, or the sometimes poorer data in areas where changes in curriculum had been introduced. Not giving a rating, or giving a Bronze rating, to a course that was new or undergoing significant change, did not feel to some panel members like a fair outcome.
45. The costing of the exercise, particularly in terms of a scaled-up panel assessment process, needs to be thoroughly assessed ahead of full implementation. A full sector subject-level TEF would require a panel greater in size by an order of magnitude and would require a comprehensive recruitment exercise.

Table 2: Main Panel members

Chair	
Prof Janice Kay	Provost and Senior Deputy Vice-Chancellor, University of Exeter
Deputy Chair	
Prof Helen Higson	Provost and Deputy Vice-Chancellor, Aston University
Members of Main Panel	
Prof Kathleen Armour	Pro Vice-Chancellor (Education), University of Birmingham
Prof Patrick Bailey	Deputy Vice-Chancellor, London South Bank University
Prof Paul Bartholomew	Pro Vice-Chancellor (Education), University of Ulster
Prof Amanda Chetwynd	Provost, University of Lancaster
Prof Peter Elias (employment expert)	Professor at Institute for Employment Research, University of Warwick
Mark Harris	Dean of Higher Education, Stockport College
Prof Christopher Hughes	Pro Vice-Chancellor (Education), University of Warwick
Prof Becky Huxley-Binns	Vice-Provost, Academic Enhancement, The University of Law Limited
Prof Mary Malcolm	Deputy Vice-Chancellor (Academic), University of Bedfordshire
Prof Steven McIntosh (employment expert)	Professor of Economics, University of Sheffield
Ross Renton (widening participation expert)	Pro Vice-Chancellor, University of Worcester
Prof Liz Thomas (widening participation expert)	Professor of Higher Education, Edge Hill University
Robin Webber-Jones	Head of Higher Education, Derby College

Main Panel members who are also subject panel chairs and deputy chairs	Subject panel role
Prof Julia Clarke	Pro Vice-Chancellor (Business and Law), Manchester Metropolitan University Chair – Business and Law

Lewis Cleminson	Vice-President (Education), Southampton Solent Students' Union	Deputy Chair – Engineering and Technology
Peter Cowan	Vice-President, The Open University Students' Association	Deputy Chair – Humanities
Diarmuid Cowan	Students' Union President, Heriot-Watt University	Deputy Chair – Social Sciences
Josh Gulrajani	Former Students' Union Vice-President (Education), University of Essex	Deputy Chair – Business and Law
Prof Carol Hall	Director of Undergraduate Education, School of Health Sciences, University of Nottingham	Chair – Medical and Health Sciences
Prof Nick Lieven	Professor of Aircraft Dynamics, University of Bristol	Chair – Engineering and Technology
Martha Longdon	Students' Union President, Nottingham Trent University	Deputy Chair – Natural Sciences
Aaron Lowman	Former Students' Union Vice-President, Brunel University London	Deputy Chair – Medical and Health Sciences
Prof April McMahon	Deputy Vice-Chancellor (Education), University of Kent	Chair – Humanities
Prof Susan Orr	Dean of Learning and Teaching Enhancement and Professor of Creative Practice Pedagogy, University of the Arts London	Chair – Arts
James Perkins	Former Vice-President (Education), City University London Students' Union	Deputy Chair – Arts
Prof Sue Rigby	Vice-Chancellor, Bath Spa University	Chair – Natural Sciences
Prof Neil Ward	Deputy Vice-Chancellor and Pro Vice-Chancellor (Academic Affairs), University of East Anglia	Chair – Social Sciences

Synthesis of findings

This section draws together OfS's findings from all the strands of evaluation in order to highlight areas of consensus, key issues and potential responses that emerged across the evidence base. It includes consideration of feedback from panels, providers and student representatives alongside quantitative evidence derived from an analysis of the metrics and indicative ratings generated, and cost data submitted by participating providers.

46. Across both models we received and considered 727 subject and provider-level cases. The exercise demonstrated that subject-level assessments and decisions about the ratings can be successfully made based on the framework currently applied in provider-level TEF. It emphasised the importance of an assessment framework that encompasses an independent panel, applying its professional knowledge and expertise to consider contextual information, provider metrics and provider submissions to reach a holistic judgement.
47. However, a strong view emerged that neither proposed model worked as constructed, and a new model should be tested that draws on the best elements identified. Significant cross-cutting issues will also need to be addressed before full roll-out. Concerns about model design were reflected in the chair's report and across the subject chair reports. Similarly the provider survey showed that, after the submissions process, providers did not support either model as proposed. Where support was expressed it was for particular features of the model only, or was on the basis of significant qualification and there was no strong preference for one model over another.

How the models generated subject-level ratings

48. Models A and B both have features that were intended to produce robust ratings while reducing the burden of TEF at subject-level. However, these features were found to introduce layers of complexity and did not produce robust rating for all subjects: the 'by exception' approach to assessment in Model A; and the application of subject-group submissions in Model B.

Model A

49. Panel members and providers were in agreement that the approach to subject submissions, where they were made, was one of the best features of Model A. The panels valued the clarity and accessibility of a five-page submission format that focused on one subject:

'The panel overall found the Model A submission to be more focused and more useful.'

– Natural Sciences Panel report

'The panel was unanimous in its preference for the single subject-level statements and considered that this approach worked best for all providers.'

– Business and Law Panel report

50. This view was supported by providers via the survey, who also favoured the provider-level submission format of Model A:

‘In reference to the writing of the submissions, the Model A approach seemed much more straight forward with a comprehensive provider-level submission of all 10 criteria and then five pages for each subject to make their case.’

– Provider comment

‘[S]ingle subject submissions were more straightforward to write in Model A.’

– Provider comment

51. However, there were concerns about the key design feature of Model A: its exception-based approach to subject-level assessment. These concerns related to both the rule for generating exceptions, and the overall concept of a by-exception model that does not provide a comprehensive assessment of all provision. In the provider survey, respondents commented on the complexity of the Model A process for generating exceptions, and there was a mixed view among providers on whether the process identified subjects that were known to have a stronger or weaker performance than the provider overall. There was significant concern that a model that does not fully assess all subjects is therefore not entirely representative. Participating providers found this frustrating as the model prevented them from highlighting and reflecting on significant areas of provision:

‘We do not believe that our provision was accurately reflected as we did not submit assessment for every subject. Based on our Model B submissions, we believe that some of our subject areas that are Silver and therefore non-exceptions have some excellent examples of best practice and would be worthy of moving up the rankings to Gold.’

– Provider comment

‘We were unable to accurately represent our provision because a large chunk of it was not an exception subject and therefore was missing from the submission.’

– Provider comment

52. These views were echoed by the panels:

‘Model A was felt to be lacking in its provision of information to applicants, and potentially misleading, with concerns around the unknown element of non-exception subjects that would inherit the provider-level rating with little interrogation. One of the key benefits of TEF to date, and future benefits of subject-level TEF, is increasing the focus on teaching enhancement and panels felt that this would not occur across the board if only a sample of subjects were fully assessed. Non-exception subjects would not have an associated submission, making it difficult for students to make fully informed choices about a subject across institutions.’

– Chair’s report

53. From the outset a key question surrounding Model A was the validity of its central premise: that the provider-level rating produced by this model would reflect teaching quality and student outcomes in most parts of the provider’s subject provision. If this premise is invalid, there is the potential for ‘non-exception’ subjects to automatically inherit a provider rating that does not

reflect the actual subject-level provision. In these cases, the subject's teaching and student outcomes would not be accurately represented by the TEF descriptor of the inherited rating.

54. Following the completion of the Model A provider-level assessments, we designed and conducted an exercise in which we asked a subset of the Main Panel to review an allocation of providers' entire metrics (i.e. the full set of provider and subject-level metrics for each provider). Reviewers reported back on how reasonable it would be for each subject within a provider to inherit the final provider-level rating rather than being fully assessed by the relevant subject panel. Each provider was reviewed twice, and we then examined the level of consensus between the two reviewers for subjects in scenarios where:
 - a. The final provider rating remained consistent with the provider initial hypothesis.
 - b. The final provider rating moved up one rating from the provider initial hypothesis (note that in Model A this was the only type of movement seen – there were no negative movements or upwards movements of more than one rating).
55. Across these combinations we found low levels of consensus that it would be reasonable for a subject to inherit the provider rating. For example, when the final provider rating remained consistent with the provider initial hypothesis, there was consensus for only 52 per cent of non-exceptions that it was reasonable for those subjects to inherit the provider rating without full assessment at subject-level. This dropped to 20 per cent where the provider final rating moved up from the initial hypothesis.
56. We also fully assessed a sample of subjects in Model A that were not generated as exceptions, which allowed us to directly compare ratings inherited from the provider outcome with ratings generated through panel assessment. Figure 4 shows that, for the sampled non-exception subjects (i.e. those subjects that were assessed but were not true exceptions), almost 40 per cent would have received a different rating if they had inherited the provider rating than if they were assessed by a panel.
57. This evidence is further supported by the subject ratings generated in Model B and their relationship with the final provider-level rating. As Figure 5 shows, approximately 30 per cent of subjects that would have been classified as non-exceptions in Model A received a rating that was different to the final provider-level rating. That is, using Model B ratings to simulate Model A outcomes also indicates that a significant proportion of provider-inherited ratings for non-exception subjects would not represent provision as accurately as full assessment at subject-level.
58. It is clear, based on the strong view of the panel and the evidence from the ratings generated, that provider metrics alone cannot provide a robust mechanism for identifying which subjects should inherit the provider rating. The importance of holistic judgment should take precedence over this design feature.

Figure 4: Model A – Subject’s rating difference from provider’s final rating, by ‘Type of exception’, all subjects

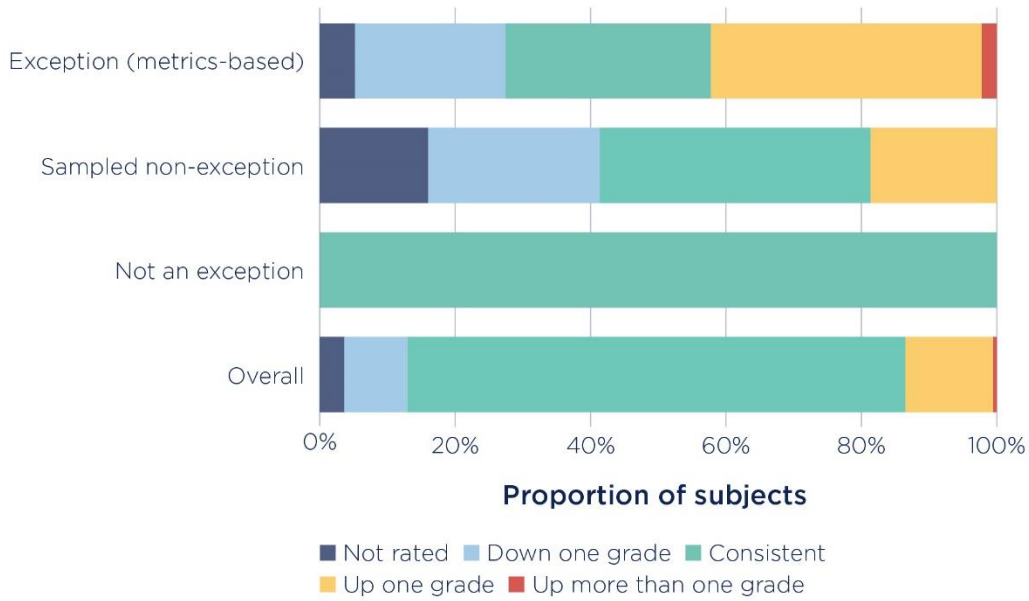
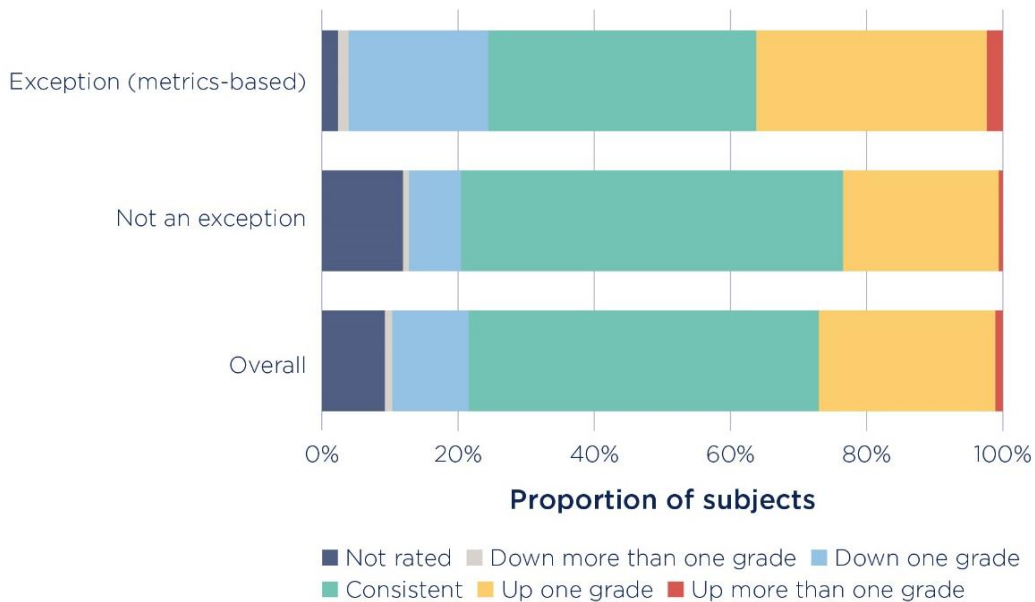


Figure 5: Model B – Subject’s rating difference from provider’s final rating, by ‘Type of exception’, all subjects



Model B

59. The key advantage providers and panel members identified in Model B was the potential to address all subjects offered by the provider. However, providers and panel members broadly

agreed that the current format of provider and subject-level submissions in Model B did not allow for the totality of an institution's provision to be represented or assessed.

'There was insufficient space to do justice to any subject (staff felt very strongly about this). This led to discussion about which subjects to prioritise and which to sacrifice – hardly ideal. Trying to knit several subjects into one qualitative submission so that assessors can pull it apart again seems mad to us!'

– Provider comment

'It provided more scope than Model A in that every subject was covered. However, the constraints in place by the subject groups made it difficult to fully and therefore accurately represent each subject and forced us to instead focus more attention on particular subjects. We would prefer a focus on single subject submissions for every subject.'

– Provider comment

'The fullness of assessment in Model B made it a more comprehensive and reliable source of information for applicants, and while the subject group submission and the subject-based initial hypothesis were weaknesses in Model B's design, overall panel members saw more opportunities for tweaks to improve this model.'

– Chair's report

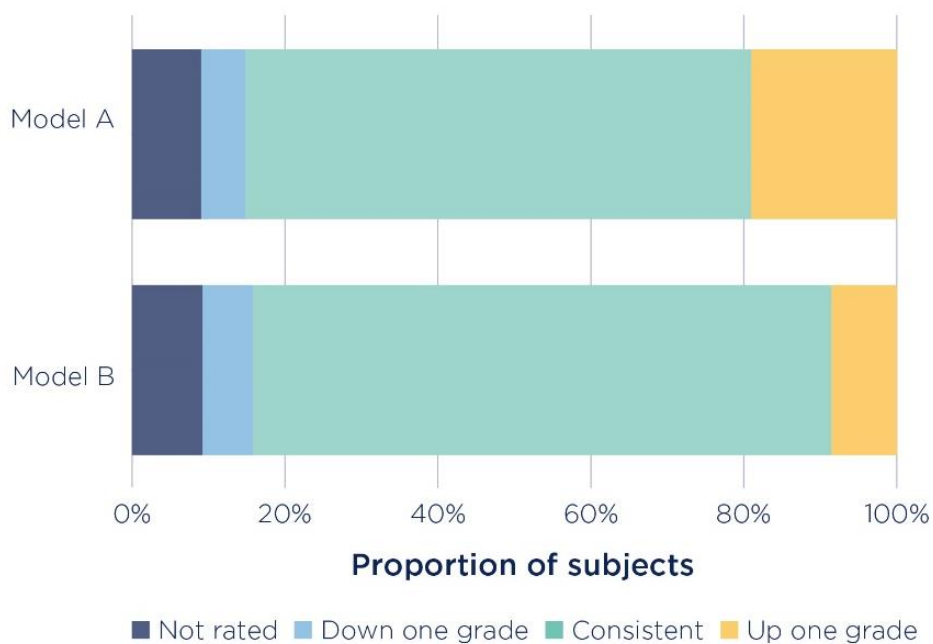
'Under Model B, assessors had to search through subject group submissions to identify material relating to the subject area under consideration, which was awkward and time-consuming.'

– Social Sciences Panel report

60. The subject-group submission format was also challenged by providers in the cost survey, where written comments suggested that workload actually increased based on the amount of editing and condensing required on the submission. Further concerns were raised by students – who saw less evidence of student engagement in Model B submissions – and by widening participation experts, who similarly reported that widening participation concerns were less visible in Model B submissions.

61. There was consensus across the panels that Model A subject submissions, where they were made, provided a foundation for better subject-level assessment than the subject-group submission format in Model B. This is reflected in the fact that they appear to have played a greater role in decision-making than in Model B – a higher proportion of assessed subjects received ratings that were higher than the initial hypothesis in Model A than in Model B, as shown in Figure 6.

Figure 6: Difference in subject panel rating from the initial hypothesis (all assessed subjects in both models)



Provider-level assessment and its relationship with subject-level ratings

62. The Main Panel’s task was more straightforward in Model A. It was analogous to the provider-level Year Three exercise and subject ratings were not part of the provider-level assessment framework. Model B led to more complex and lengthy discussions, as was expected given the need to consider subject-level ratings through the subject-based initial hypothesis (SBIH), and a more limited provider-level submission.
63. The panel undertook a dedicated calibration exercise to test the presentation of SBIH data, and panel members undertook separate calibrations to test the influence of using the SBIH at different points in the process – this led to a common approach for the actual assessment whereby consideration of the SBIH represented the first assessment step. Overall, the SBIH was found to cause anchoring and to compound issues of ‘metrics capture’ (which refers to panel members being reluctant to move their holistic judgement away from the position indicated by the metrics alone, discussed at paragraph 114) and ‘silverness’ (which refers to the higher rate of Silver initial hypotheses observed at subject-level, compared with provider-level, as described at paragraph 94).
64. As shown in Figure 7, twice the number of Bronze ratings were arrived at in Model B than in Model A, for providers taking part in both models. Furthermore, some providers taking part in both models moved down one grade from the step 1a initial hypothesis in Model B, but not in Model A (Figure 8). This could relate to the anchoring issues surrounding the SBIH identified by the panel.
65. Panel members and providers did not support the algorithmic format of the SBIH. However, it was felt that a profile of subject ratings was useful contextual information for the provider-level

rating. Panel members recommended that this could include some systematic detail around the panel outcomes, for example how panel ratings compared with the initial hypothesis.

66. There was a general sense that the processes needed to be sequenced for practical purposes, but it was felt that the concept of provider-level assessment coming either first (Model A) or second (Model B) was less important than framing more coherent and focused criteria (paragraphs 108-109). Panels were content that judgements at provider and subject-level did not need to always inform each other and that it was right that there could be differences. Some panel members felt there should be a form of relationship between provider and subject-level assessment, but not one that is formulaic.

‘There was support within the Main Panel for keeping subject-level and provider-level submissions, but decoupling them, potentially extending to using different criteria to judge them. This arose from a majority of Main Panel members rejecting the SBIH and the heavy influence it was felt to have on provider-level judgements.’

– Main Panel Chair’s report

Figure 7: Profile of provider-level final ratings by model, for providers in both Model A and Model B

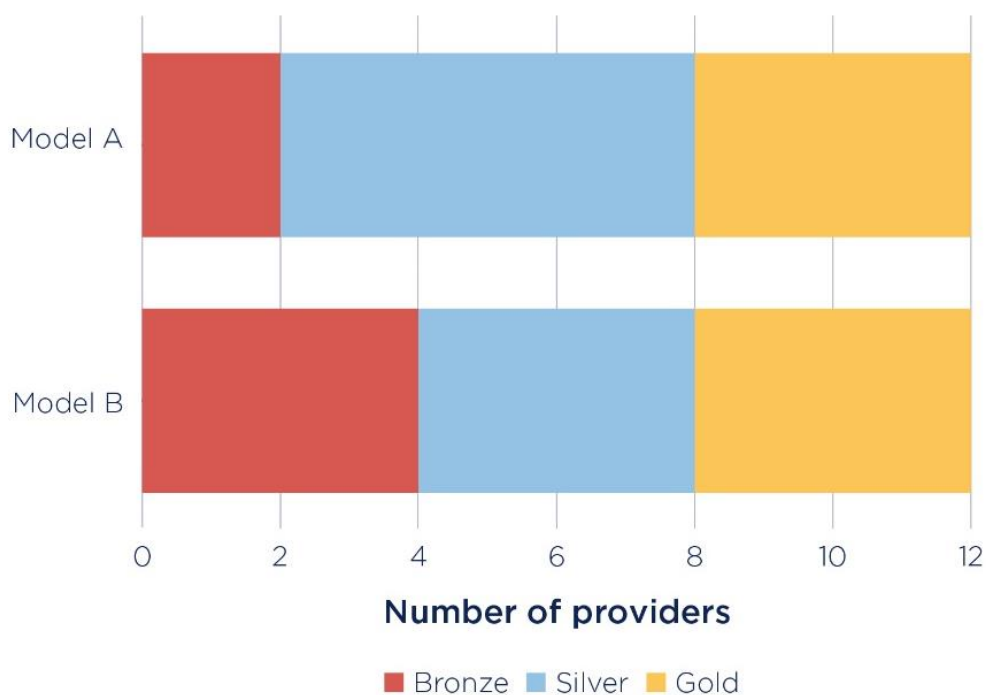
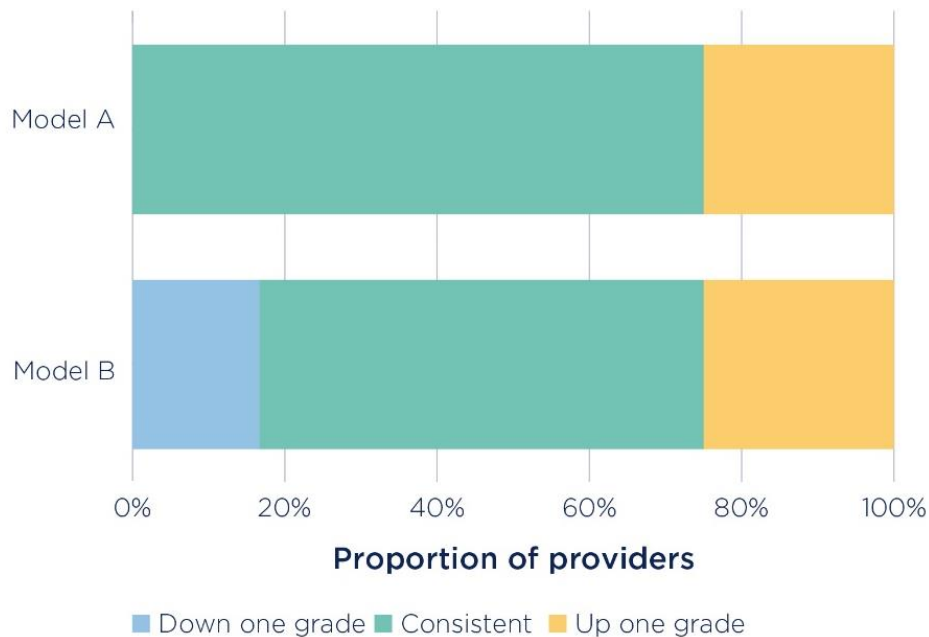


Figure 8: Provider’s rating difference from provider’s 1a initial hypothesis by model, for providers in both A and B



A more comprehensive model

67. There is broad support for a more comprehensive model of assessment for generating ratings at provider and subject-levels from participating providers, the pilot panels and from students.
68. At our student focus groups, there was a consensus among participants that Model A produces outcomes which are less useful to applicants. For example, a final provider-level rating may well differ from its initial hypothesis, and students would be unaware of how a non-exception subject ‘sits’ in the context of its inherited provider-level rating. The focus groups felt strongly that a model which fully assesses each subject would produce a more meaningful and accurate set of ratings for applicants and students. This view was supported by student panel members and student representatives at provider briefing events.
69. The specific idea of assessing all subjects in full was also independently proposed by providers who took part in both models.

‘Having a submission from the provider and each subject, all covering all 10 criteria is the fairest and simplest way of having TEF at subject-level.’

– Provider comment

‘We believe this approach is the only way to take subject-level TEF forward in a representative and accurate format as every area would be individually represented as opposed to over-reliance on particular areas (either the provider submission in Model A or Group Submissions in Model B). This format would also allow us to better embed TEF

within the institution as it would become business as usual if every subject was reviewing their subject on an equal footing.’

– Provider comment

‘[I]f the aim of subject TEF is to aid student choice, submitting all subjects will ensure fairness and transparency because there will be a comparable amount of information available for students interested in each subject. This approach is also fairer and more transparent for the institutions.’

– Provider comment

‘Having separate subject submissions instead of a group submission would have been less time consuming. [...] The least time-consuming subject-level TEF would be one that required a submission from each subject (e.g. five pages) and a provider-level submission (e.g. 15 pages across all criteria).’

– Provider comment (cost survey)

70. We tested this proposal more widely in the formal provider survey. Crucially, 10 out of 12 providers that participated in both models, including all four further education colleges and alternative providers, agreed or strongly agreed with the proposal. Of the wider group of pilot participants, 50 per cent supported the proposal while ten percent remained neutral (see Annex C, paragraph 34).

71. Among those who agreed with the proposal, it was commonly suggested that it was easier from a provider perspective to manage the process to be inclusive of all of provision. Providers also reflected that it allowed a better and more meaningful representation of their provision, which was positive for institutional engagement and enhancement as well as the quality of information that would be generated for students. Providers that took part in both models reported that Model B would in fact be less burdensome if separate submissions were made for all subjects.

72. There was universal support for a comprehensive model from the panels, where it was clearly articulated that it made more sense to first make the process robust, and then focus on how to make it light-touch, rather than the other way round:

‘The second year of the subject pilots should be taken as an opportunity to test refinements to a single refined model which draws upon the best elements of both models.’

– Chair’s report

‘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.’

– Natural Sciences Panel report

‘A [...] more popular option is a version of Model B for all subjects with no grouped subject submissions.’

– Arts Panel report

73. The OfS has since produced an estimation of the full sector cost of a refined model, a subject-level exercise that is similar in structure to Model A but with all populated subjects assessed. The findings of this cost estimate are given at Annex D.

Rating structure and descriptors at subject-level

74. Subject panels reflected dissatisfaction with using the existing rating structure at subject-level. This was the result of a combination of factors which arise at subject-level due to smaller cohort sizes and ‘metrics capture’ (paragraph 112). Subject panels found that the ratings Bronze, Silver and Gold applied to too large a range of performance at subject-level. In particular the gap between ‘low’ and ‘high’ Silvers was problematic, and borderline ratings required significant debate to resolve. These issues were compounded by the associated ratings descriptors, which were not tailored enough at subject-level. The subject panels therefore felt less confident moving away from the initial hypothesis on the basis of the ratings descriptors in their current form:

‘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.’

– Natural Sciences Panel report

‘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.’

– Social Sciences Panel report

75. The subject panels recommended that refinement of these descriptors could go some way towards resolving the difficulties experienced in subject-level assessment.

76. When the ratings structure and descriptors were discussed at the Main Panel, it was acknowledged that the independent review is expected to consider the number and names of ratings. The Main Panel agreed that a constructive commentary or statement of findings should form part of the outcomes at both provider and subject-levels (and such qualitative feedback could mitigate the issue of borderline ratings mentioned in paragraph 73). There was support for ratings that profiled different aspects of quality at provider-level, including a profile of subject ratings. It was also recommended that the rating descriptors be amended in line with differentiated criteria at subject and provider-level.

Cross-cutting issues

Student engagement

77. The professionalism, expertise and insight of student panel members were commended by the panels. They played a full and equal role in the assessment and provided constructive challenge from the student perspective. In addition, the OfS sought to gain student views on developing student engagement with subject-level TEF.
78. The OfS collected student-specific feedback from panel members and from student representatives more widely via:
- a report drafted by the subject panel deputy chairs
 - a focus group with student panel members
 - a focus group with student representatives from providers participating in the pilot
 - two focus groups with student representatives who did not participate in the pilot.
79. Panel members and student representatives identified a range of options for the OfS to consider to maximise student roles in the assessment process and also to strengthen student engagement in provider submissions. For example, increasing the proportion of student panel members across the main and subject panels, and improvement of student input in submissions. It was suggested that the current student engagement criterion could be split into two distinct aspects: student engagement with learning, and student partnership. There was also consensus that the metrics should be as relevant as possible to what matters most to students, and additional National Student Survey (NSS)-based metrics could be included to address the importance of learning resources and student partnership with high education providers. Annex A sets out the student views collected in further detail.

Assessability of subjects

80. As recognised both in the pilot specification and through the DfE technical consultation, a key challenge in subject-level assessment relates to limitations in the data when disaggregated at subject-level. Two known issues were reviewed in the pilot:
- a. **Reportability:** Some subjects do not have a full set of reportable metrics. This may be due to very small cohort sizes, new subjects that do not yet have reportable data, particularly employment metrics, or variations in survey response rates. While a provider may have a full set of metrics at provider-level, a number of its subjects may have at least one core dataset missing.
 - b. **Assessability:** Some subjects have very low cohort numbers and therefore confidence in the statistical reliability of the data is low. Such cases generate few or no flags, and therefore most cases are Silver 'by default' at the step 1a initial hypothesis (paragraph 93).
81. To explore these issues, the pilot sample was deliberately constructed to include a range of examples, from subjects with very large cohorts and complete sets of metrics through to contrasting examples where the subjects combined very small cohorts with multiple non-

reportable metrics. The panels found that at the extreme end they were unable to use the metrics to inform the assessments.

Scope issues

82. In addition, when metrics were first released at the start of the pilot, some providers advised that some of their subject workbooks covered provision that was potentially not appropriate to assess. In discussion with providers, a small number of these were removed from the pilot in advance; these were typically subjects which had been closed⁹, where metrics were only populated by legacy cohorts and as a consequence often generated multiple metrics with non-reportable data.
83. During the pilot we found further examples of cases that may be inappropriate to assess: 'subjects' in which no whole programme is taught. This included metrics that were generated on intercalated degrees and foundation year courses. In many cases, foundation year course provision is franchised out to another provider and when returning to their registering provider students may study a different subject for the remainder of their degree. In the pilot and in previous TEF exercises, the assignment of a teaching provider for the purpose of TEF metrics is calculated on the basis of where the student is taught in the majority in their first year of study. The associated outcomes for all these students would be assigned to the provider that offered the foundation year, despite their outcomes being likely to be more reflective of the registering provider rather than the franchised provider. Franchised providers fed back that it was difficult to address any activity beyond the foundation year they provided when writing their submissions. It was also challenging for the subject panels to assess such cases.
84. These findings highlight the need to develop explicit criteria for determining whether a subject should be 'in scope' for subject-level assessment, and additional processes would need to be built into subject-level TEF to confirm which subjects at each provider are in scope for assessment. This may also carry implications for metrics definition, which OfS is considering

Data limitations affecting assessment

85. As anticipated, panels found it challenging to make use of workbooks that contained non-reportable metrics, to inform their judgements. Both providers and panels raised a specific issue about subjects that consist of entirely new courses, which do not yet have graduating cohorts and hence no Destinations of Leavers from Higher Education (DLHE) or Longitudinal Employment Outcomes (LEO) data – or any alternative way of demonstrating employment outcomes in their submissions. It seemed illogical for a course with no graduating cohort to need to demonstrate impact against the Student Outcomes and Learning Gain (SO) criteria. The subject could well receive a lower than deserved rating due to this lack of data, despite having wholly positive evidence in the other aspects: panels expressed concern that this could be a disincentive against introducing new and innovative courses. Some panels recommended that such subjects could receive 'provisional' ratings. The OfS will test ways of handling new and emerging provision in the second year of subject pilots.

⁹ A subject is considered to have been closed when a provider is no longer enrolling new undergraduate students onto any of the courses that it had previously offered in that subject, has no current intention to recruit students onto any such courses in the future and ceases to advertise that subject in its undergraduate course offering.

86. The subject panels also found it difficult to tether their judgements where student cohorts were small. Subjects with small cohorts were often found to have submissions that contained limited evidence, increasing concerns around rating robustness.
87. These issues relating to scope, non-reportable metrics and cohort sizes came to the fore during the assessment of Model B subjects, which were assessed prior to subject-level assessment in Model A. The panels attempted to arrive at a rating for all subjects in Model B but found that there were some subjects where the metrics – in combination with the submissions – contained insufficient evidence to make a best fit judgement about the rating.

'No rating' outcomes in the pilot

88. We undertook a moderation exercise with the Main Panel for Model B cases where the subject panels found reaching a judgement problematic based on data limitations. The panel concluded that it was not appropriate to force a rating for a subject where there was insufficient evidence to make a judgement, and in particular it would not be appropriate to simply default to a Silver rating. Applying the provider rating in this context was not seen as robust and led to other unintended consequences.
89. During the subsequent Model A subject assessment phase, on the advice of the Main Panel, subject panels were given a more systematic option of arriving at 'No rating' where they judged there was insufficient evidence available in the metrics and the submission.
90. A summary of the pilot ratings demonstrates that the number of reportable metrics, as well as cohort size, affected whether there was sufficient data to inform an assessment. Of the 210 subjects the panels assessed in Model A, they considered two to be out of scope (even though they had all data sources and sizeable populations), and they were unable to rate a further 18, of which:
- 13 had more than one missing data source
 - 14 had student populations of less than 40.
91. Eight subjects with a population of less than 40 received a rating: all of them maintained the 'default' Silver position. Overall, no subject with a population of less than 40 was rated either Gold or Bronze.
92. Figures 9 and 10 demonstrate that the fewer reportable core metric sources, the higher the likelihood of a 'No rating'. As Figure 9 shows, over 80 per cent of cases with only one missing data source received a rating. When two data sources are missing, this drops significantly to 50 per cent.

Figure 9: Model A – Movement in subject rating from initial hypothesis, by ‘Number of non-reportable data sources (max 3)’, panel-assessed subjects only

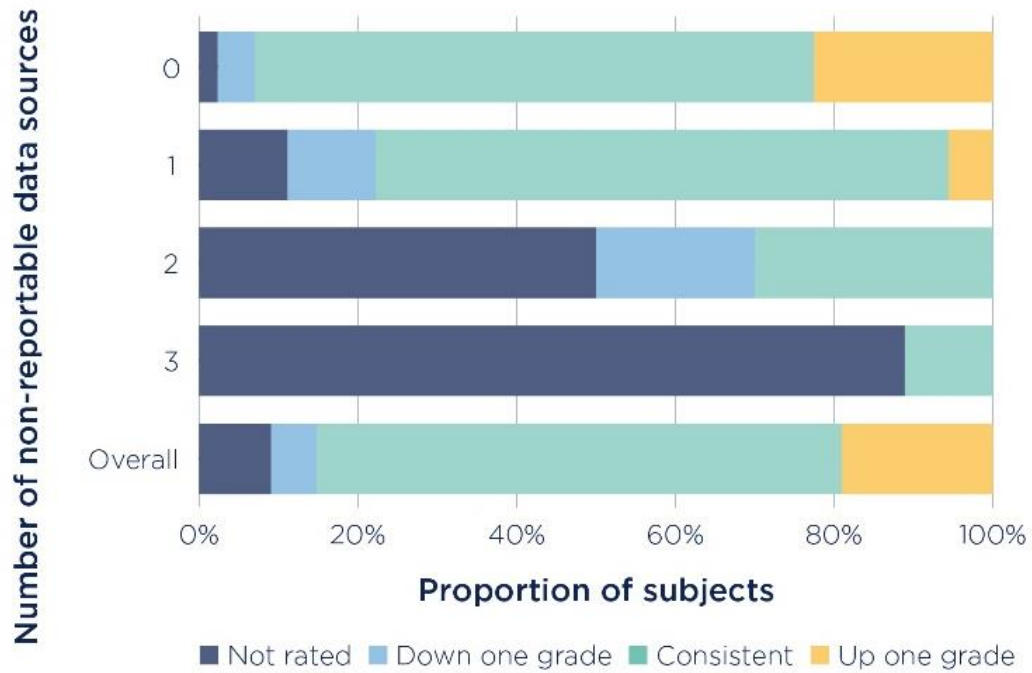
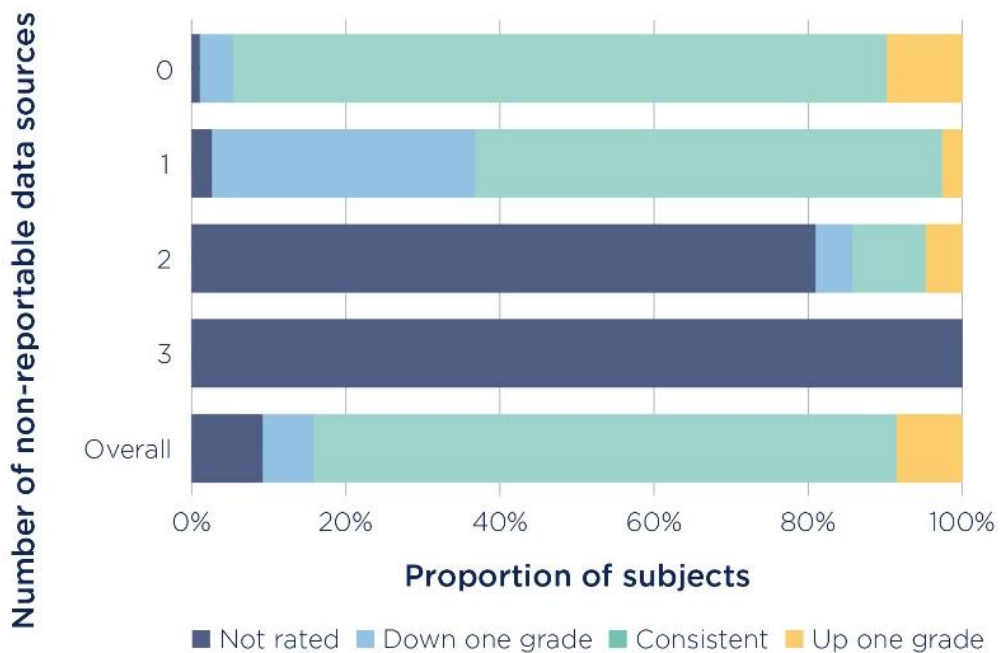


Figure 10: Model B – Movement in subject rating from initial hypothesis, by ‘Number of non-reportable data sources (max 3)’, all subjects



93. All panels independently came to the view that a cohort size threshold for assessment would be necessary, to enable the assessments to be meaningfully informed by metrics. While some made suggestions about what a suitable threshold might be, the OfS undertook further analysis to help inform consideration of a cohort threshold.

‘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 [it is] not Bronze would be potentially unfair.’

– Business and Law Panel report

‘Lowered levels of confidence associated with small cohorts leads 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.’

– Arts Panel report

‘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.’

– Social Sciences Panel report

‘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.’

– Business and Law Panel report

Additional OfS analysis

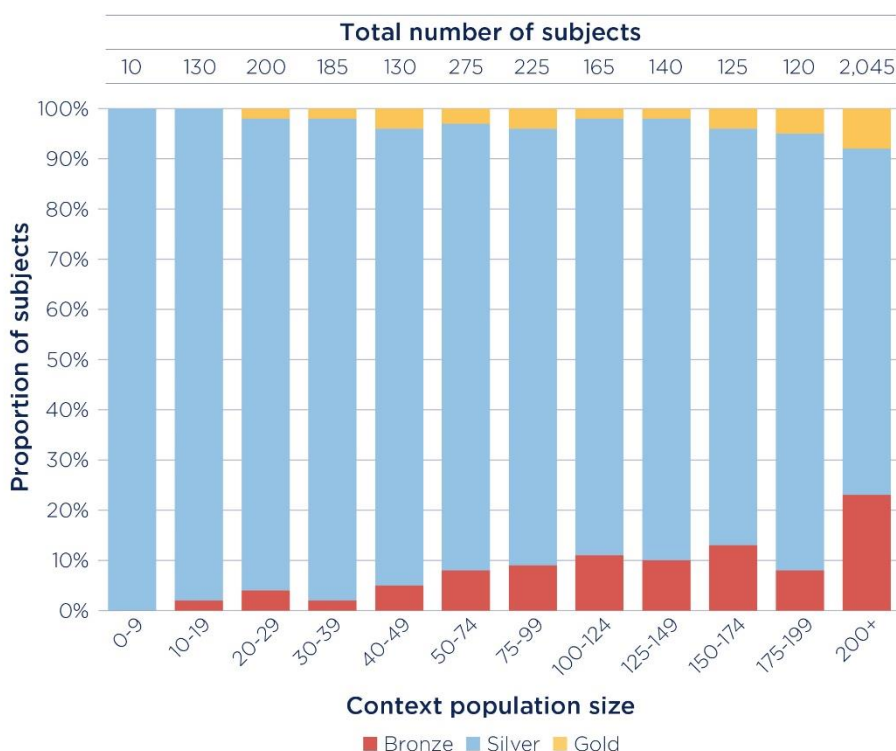
94. In addition to finding some subjects with insufficient data to inform assessments, panels were concerned more generally by the number of subjects that ‘defaulted’ to Silver in the initial hypothesis at step 1a. This ‘silverness’ at subject-level is caused by the neutralising effect of fewer flags when there are smaller cohorts and fewer reportable data sources. The OfS undertook additional analysis of this issue and Figure 11 shows the step 1a initial hypotheses for subjects with different population sizes (based on the contextual data)¹⁰. It **excludes** subjects with more than one missing data source.

¹⁰ The contextual population is the annual average – over the most recent three years – of all undergraduate students in all years of study, in majority mode. So, for example, if there have been 10 entrants every year and all entrants complete a three-year programme, the contextual population would be 30. Figure 11 includes all providers with suitable metrics at provider-level, and at least two reportable data sources.

95. In summary:

- a. With population sizes up to 40, the vast majority of subjects ‘default’ to Silver due to a lack of flags and there appears to be very limited discriminatory value in the data.
- b. With populations between 40 and 200, there is some but relatively limited discriminatory value in the data.
- c. With population sizes of 200 or more, there is a more reasonable spread across the three rating categories.

Figure 11: Proportion of subjects with each step 1a initial hypothesis, by subject’s contextual population size



96. The need for thresholds raises questions about how subjects falling below the thresholds should feature in subject-level TEF, and the impact they would have on the coverage of subject TEF ratings. The OfS has undertaken an analysis to test how setting a minimum threshold would affect coverage. The analysis encompasses subjects with at least two reportable data sources. Table 3 shows that setting a threshold at a level that starts to provide discriminatory value (at least 40 students) would have a negative impact on apparent subject coverage, excluding up to 28.4 per cent of subjects where all providers are taken into account, but would have minimal impact on the coverage of students.

Table 3: Coverage estimates for all providers (total number of students 1,780,000)

Thresholds	Number of students covered	% of subjects covered	% of students covered
At least 2 reportable data sources, any student numbers	1,743,000	84.4%	97.9%
At least 10 students and 2 reportable data sources	1,743,000	84.0%	97.9%
At least 20 students and 2 reportable data sources	1,740,000	80.2%	97.8%
At least 30 students and 2 reportable data sources	1,733,000	75.3%	97.4%
At least 40 students and 2 reportable data sources	1,727,000	71.6%	97.0%
At least 50 students and 2 reportable data sources	1,718,000	67.9%	96.5%

97. The greatest impact of a minimum threshold would be on providers that teach across a range of subjects with low student numbers – most typically further education colleges. This is illustrated in Table 4.

Table 4: Coverage estimates for all further education colleges (total number of students 95,000)

Thresholds	Number of students covered	% of subjects covered	% of students covered
At least 2 reportable data sources, any student numbers	79,000	69.7%	83.8%
At least 10 students and 2 reportable data sources	79,000	68.6%	83.6%
At least 20 students and 2 reportable data sources	77,000	58.3%	80.8%
At least 30 students and 2 reportable data sources	72,000	46.4%	75.6%
At least 40 students and 2 reportable data sources	66,000	37.7%	62.3%
At least 50 students and 2 reportable data sources	60,000	29.5%	63.4%

98. For alternative providers, the main coverage limitation relates to non-reportable metrics, a situation that is expected to improve significantly over the next few years. Increasing the size threshold has limited additional impact on alternative providers, as shown in Table 5.

Table 5: Coverage estimates for all alternative providers (total number of students 45,000)

Thresholds	Number of students covered	% of subjects covered	% of students covered
At least 2 reportable data sources, any student numbers	42,000	74.2%	94.4%
At least 10 students and 2 reportable data sources	42,000	73.7%	94.3%
At least 20 students and 2 reportable data sources	42,000	72.2%	94.2%
At least 30 students and 2 reportable data sources	42,000	67.0%	93.6%
At least 40 students and 2 reportable data sources	42,000	63.6%	93.0%
At least 50 students and 2 reportable data sources	41,000	58.9%	91.9%

99. The panels agreed that ‘No rating’ was unhelpful terminology and an unsatisfactory solution, and that it would be appropriate to consider different options based on the cause of the data limitations. For example it may be appropriate to consider provisional (or similar) awards for new courses or subjects with small cohorts, but no awards for subjects that do not offer whole courses or that have been completely taught out.
100. It was recommended that further work should be done to explore these options and how they would be interpreted by applicants, and the impact of thresholds on coverage. It was noted that the potential impacts may not be limited to small providers – there may also be larger providers with provision comprising small subject cohorts.

Subject categorisation

101. The Common Aggregation Hierarchy level 2 (CAH2) subject classification system was found to be at broadly the right level of aggregation for producing robust ratings at subject-level, and CAH2 compared favourably with other potential classification systems¹¹. Most panels and participating providers identified some refinements that could be made to CAH2 to better reflect the diversity of provision in UK higher education. However, the issue of adequately representing interdisciplinary subjects which map to two or more CAH2 categories will require further consideration. The CAH2 subject groupings will also need to be made clear when communicating ratings to applicants and current students.

Contextual data

102. It was observed in both models that the process would have been much smoother if there was an enhanced set of contextual data at subject-level. Data relating to the overall number of students for each subject was supplied to providers and panel members, but data about individual courses was not available. Because of the aggregation of courses to subject-level, it was unclear which specific courses were covered in the metrics unless they were listed by the provider in the written submission (which was not commonly done). This lack of clarity added to the complexity of assessment, particularly when subjects encompassed interdisciplinary provision or if only a subset of courses within a subject might be expected to be accredited.
103. These issues were exacerbated further in Model B, where it was not only difficult to disentangle courses, but very often difficult to disentangle subjects as well. It was not efficient for panels to spend time in either model on investigating the structure of provision.
104. Additionally, it was clear that subject submissions and assessments would have been aided by setting each subject in the overall context of the provider’s mission and structure.

‘The fact that the panel for each subject-level did not receive the provider submission has meant that we have had to replicate “generic” topics across every submission to make each submission as a standalone document. This has meant that we have had fewer pages to actually address local provisions. The purpose of the submissions was to: add additional

¹¹ DfE research, undertaken by IFF research, has shown that CAH2 is the most usable and preferred classification system for applicants when compared with CAH1 (23 subject areas) and a broader classification containing seven subject areas (<https://www.gov.uk/government/publications/teaching-excellence-framework-and-informing-student-choice>).

context to standard data, address the performance against the metrics, put forward evidence against the assessment criteria and explore performance of specific student groups against the split metrics and we feel that the page lengths for the submissions were not sufficient.'

- Provider comment

'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.'

- Social Sciences Panel report

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

- Engineering and Technology Panel report

Accreditation

105. In general the panels did not feel that a compulsory declaration of accreditation should be a requirement. As the significance of accreditation varies widely across subject areas, developing a standardised approach was not viewed as useful. Some panels reflected that further guidance on referencing accreditation in submissions could be given to providers:

'The panel would have liked to have seen stronger and more detailed coverage of PSRB accreditation issues where these applied.'

- Social Sciences Panel report

106. PSRB and employer representatives on the Medical and Health Sciences Panel highlighted the importance and potential risks of subject TEF's relationship with professional accreditation, and recommended that further work could be done in this area:

'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.'

- Medical and Health Sciences Panel report

Interdisciplinarity

107. Providers and panel members found the articulation and interpretation of narratives for the three broad interdisciplinary subject categories challenging in a number of cases. For some providers, the diversity of courses mapped to these categories meant it was difficult to produce a submission that fully addressed very different student experiences. As noted in paragraph 100, this issue was further complicated by the subject-group format of Model B subject-level submissions.

'[T]he 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.'

– Humanities Panel report

108. It was suggested by both providers and panel members that further contextual data on individual courses and the number of full-time equivalent (FTE) students mapped to other CAH2 categories would benefit assessment across all areas of provision, and be particularly helpful in supporting panel member understanding of interdisciplinary provision:

'This would give the [panel members] an indication of the volume of interdisciplinary working, and would provide a better foundation for the authors, that doesn't require wasted space identifying [and] explaining the population.'

– Provider comment

109. It was also observed that panels could be further supported with additional interdisciplinary expertise. The OfS will test and develop new panel roles that champion interdisciplinary provision during the second pilot exercise.

Provider- and subject-level criteria

110. The pilot drew on the same set of criteria at provider-level and at subject-level in both models (though in Model B provider-level submissions were focussed on TQ2: Valuing teaching, LE1: Resources and SO3: Positive outcomes for all). While there was agreement across panels that these criteria were useful, there was also a universal panel view that it would significantly enhance the process if there was a considered differentiation between the subject- and provider-level focus, and that this needed to feed through from criteria to the ratings descriptors, submissions and panel assessment.

'There was support within the Main Panel for keeping subject-level and provider-level submissions, but decoupling them, potentially extending to using different criteria to judge subject-level and provider-level ratings.'

– Chair's report

111. Initial discussions with the Main Panel suggest that at provider-level a greater focus should be placed on the overall institutional environment and strategic oversight, while at subject-level a greater focus should be placed on pedagogy and the student experience. Consideration should be given to providing exemplars of impact mapped to each criterion (though care would need to be taken to ensure these are not framed in such a way that they drive homogenised submissions). Any reframing of the criteria should go hand in hand with considerations in relation to metrics mapping and coverage.

Assessment steps

112. The pilot has demonstrated that the current provider-level TEF method of assessment – combining metrics, submissions and holistic judgement – can be successfully applied at subject-level. From the outset the Main Panel Chair highlighted to panel members the importance of holistic academic judgement; however, there was consensus across the panels that, despite confidence in their decision-making process, the parameters did not permit as much moderation of initial hypotheses as would be desirable.
113. This was due to a combination of factors, the first of which was the prevalence of Silver initial hypotheses. The move to subject-level creates proportionally more subjects whose 1a initial hypothesis is Silver. It should be noted that a proportion of these were what the panel described as ‘technical’ Silvers – in that non-reportable metrics led to a neutral score.
114. Secondly, a set of issues led to a sense of ‘metrics capture’:
- a. Limitations in the evidence that submissions were able to provide in terms of addressing metrics performance as well as providing additional evidence against a broad set of 10 criteria. This was especially the case for the group submissions in Model B.
 - b. Complexity of the assessment design meant that guidance and training were heavily focussed on clarifying how to correctly interpret metrics for step 1. There was very limited time available for the calibration exercise, which was the main training opportunity to focus on the submissions and holistic judgement. Further development of guidance and training materials on assessing submissions and reaching a holistic judgement was recommended for the second year of pilots.
 - c. As noted at paragraph 75, ratings descriptors and outcomes could be made more focussed and meaningful at subject-level. The descriptors applied in the pilot at subject-level led to some risk aversion: while panels followed the process it was noted that in a real exercise panels would have been more concerned by the reputational consequences of following the process to a rating that could be incorrect or inappropriate. It was recommended that the OfS should test refinements to descriptors in the next subject pilot.

‘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’.

– Humanities Panel report

‘The Gold, Silver and Bronze ratings descriptors did not seem entirely appropriate for subject-level ratings and could be tweaked at both subject and provider-level’.

– Chair’s report

Teaching intensity

115. The majority of providers found it too difficult or too resource-intensive (or both) to accurately and robustly capture teaching intensity information in its current form. There was also a very

limited response to the student survey. The survey received only 4,880 full responses out of approximately 113,000 potential responses (a 4.3 per cent response rate). It should be noted that the tight timeline of the pilot meant that the survey window encompassed the winter holiday period; this will have impacted significantly on response rates.

116. Providers found the generated metrics difficult to interpret and their use in provider submissions was therefore limited – more than half of providers did not refer to teaching intensity data in their submissions at all. Again, the compressed timescale of the pilot meant that providers had limited time to translate the data into their narratives.

‘We struggled to make head or tail of them to be honest and would have needed far more training to do so.’

– Provider comment

‘We were unable to make much use of the data given that it was neither clear nor robust.’

– Provider comment

117. Similarly, no panel found it possible to meaningfully interpret and use teaching intensity data in its assessments. Therefore it was felt that no judgements of quality surrounding teaching intensity could be made, and panels reported that it did not influence the final ratings. As noted, it was not always triangulated within provider submissions, although in a couple of instances provider narratives around the gross teaching quotient (GTQ) did provide useful information on mixed teaching approaches. Panels were given standard reference statistics (for example, relating to the distribution of GTQ within a subject) but it was clear that no consensus could be reached about either threshold or comparative expectations of contact hours in relation to quality. The panels reported that GTQ in particular is difficult to interpret as it encourages a focus on quantity rather than quality of teaching, but ‘more intense’ does not necessarily mean ‘better’. These points are also reflected in the findings from the latest Higher Education Policy Institute / Advance HE Student Academic Experience Survey report, which found that ‘Gold-rated institutions [...] also have strong levels of independent study’ and that ‘class sizes tend to be higher at Gold-rated institutions’¹².

‘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’.

– Arts Panel report

118. It should also be noted that if teaching intensity was a mandatory element of subject TEF, it might represent a disproportionate cost compared with the rest of the exercise. Our estimates for both Model A and Model B indicate that returning teaching intensity data for all subjects on average comprised approximately 50 per cent of the total cost of participating in the subject

¹² Higher Education Policy Institute / Advance HE 2018 Student Academic Experience Survey. Available at www.hepi.ac.uk/2018/06/07/2018-student-academic-experience-survey/.

TEF pilot (Annex E, paragraph 15). This indicates that teaching intensity does not provide value for money relative to the wider subject-level exercise.

119. It should be recognised that contact hours remain an important issue for students. Research commissioned by the OfS, conducted by a consortium of students' unions¹³, found that 91 per cent of students they surveyed considered that the number of contact hours per week was 'very important' or 'somewhat important' in demonstrating value for money. The OfS is taking forward further work in this space, as evidenced by its 2018-19 business plan, in relation to value for money and information, transparency, and advice and guidance for students and applicants.

Grade inflation

120. In the pilot, grade inflation information was available for the provider-level assessment of providers with degree awarding powers. Panel members observed that some of the provider narratives surrounding grade inflation were of interest, but ultimately found it difficult to resolve the fact that the intended measure of grade inflation works in tension to one of TEF's key purposes: to encourage enhancement in teaching and outcomes for all students, of which attainment is an aspect. Main Panel members reported that grade inflation data had little or no impact on their holistic assessment.

'The metric (and expected commentary) is premised on the idea that improved attainment is a problem'.

– Arts Panel report

'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'.

– Medical and Health Sciences Panel report

121. It was suggested that grade inflation should be dealt with as a standards issue through regulation and other sector-led initiatives. For example, the regulatory framework¹⁴ and quality assessment arrangements will ensure that providers with degree awarding powers maintain academic standards, and the OfS business plan articulates a strategic objective to ensure that qualifications hold their value over time. The OfS is developing approaches to ensure the

¹³ 'Value for money: the student perspective', commissioned by the Office for Students and led by a consortium of students' unions, available at <https://studentsunionresearch.com/>, page 16.

¹⁴ See www.officeforstudents.org.uk/advice-and-guidance/regulation/the-regulatory-framework-for-higher-education-in-england/.

reliability of degree standards, which includes a forthcoming analysis of patterns and trends in degree classifications over time¹⁵.

122. The panels were also concerned that the incorporation of grade inflation data into assessment sent contradictory messages in relation to the pressing need to close gaps in attainment between students from advantaged and disadvantaged groups.

“[G]rade 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’.

– Humanities Panel report

‘[T]here 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’.

– Arts Panel report

123. The Main Panel recommended that supplementary data on differential degree attainment could be developed and tested at provider-level in the next subject pilot.

Widening participation

124. Two widening participation experts operated as full members of the Main Panel and had the opportunity to observe subject panels. Through their observations it was noted that many subject panels discussed widening participation considerations through subject-level assessment. Panel members were able to follow processes relating to widening participation split data. However, there was variable knowledge relating to sector widening participation issues (experience in widening participation was a desirable rather than essential criterion in recruitment for academic panel member roles). The widening participation experts advised that widening participation expertise should play a stronger and more defined role at subject-level, a recommendation that was echoed across the panels. A proposal for how to better support appropriate expertise at subject-level was suggested, where a member of each subject panel acts as a link with a small group of widening participation experts on the Main Panel.

125. Concern was raised in relation to differential attainment and outcomes. The widening participation experts highlighted that benchmarking within splits can in some cases underplay absolute differences between the split group: we know there are sector-level outcome differentials between groups with and without protected characteristics¹⁶, so it is possible for splits to be positively flagged through benchmarking but for a gap to remain in absolute terms.

¹⁵ It should be noted that the UK Standing Committee on Quality Assessment is working with other sector agencies to develop sector-recognised standards beyond the baseline threshold, which could be used to address standards at classification boundaries.

¹⁶ See www.officeforstudents.org.uk/data-and-analysis/differences-in-student-outcomes/.

The experts advised that this issue could be resolved through additional panel member training, particularly surrounding benchmarking data and what is included or omitted:

‘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’.

– Widening participation expert report

Employment metrics

126. The pilot drew on DLHE data but panel members and providers were aware that this will be replaced by the Graduate Outcomes survey (GOS) in future TEF cycles. GOS will be first carried out in September 2019, with the first data published in January 2020. In the longer term, this data source should present a more robust representation of graduate employment outcomes. However, there will be transitional issues and risks affecting TEF from Year Five (from autumn 2019), including:

- a. Whether the response target rates (70 per cent for full-time UK-domiciled undergraduates) will be widely achieved at subject and at provider-level, especially in the initial years of the survey.
- b. Whether and how to present and incorporate into assessment a single year of GOS data, and how panels then triangulate this with earlier DLHE data (which the DfE has proposed remains part of the assessment until three years’ worth of GOS data is available) and with LEO data.
- c. The need to test that benchmarking factors are robust for the new dataset and establish appropriate suppression thresholds.

127. While the second subject pilot cannot explore the use of GOS data, it was recognised that the discontinuation of DLHE will need to be addressed. To future-proof subject-level TEF in a changing data landscape, it was recommended that the OfS develop and test changes to core and supplementary employment metrics in the next pilot.

Scalability and potential impacts

128. The first subject pilot focused on testing the design features of Models A and B, and how TEF evidence, criteria and processes translate to subject-level. Ongoing evaluation will further consider issues of scalability and the potential impacts a subject-level exercise could have for higher education providers across the UK, and their provision, current students and applicants. However, all of these considerations are dependent on the final design of the subject-level TEF model.

Provider cost and burden

129. Evidence from the pilot cost study suggests that there is no statistically significant difference between the total cost to providers of the two models piloted. Regardless of which model is selected, scaling up our estimates to a sector-level estimate implies a lower bound of £5.9 million and an upper bound of £9.2 million if all English providers with 500 or more students

had participated in this year's pilot, and a lower bound of £9.1 million and upper bound of £13.8 million if all providers with eligible metrics had participated.

130. Cost survey data has been used to estimate costs to the sector of a proposed revised model as approximately £16 million for all English providers with more than 500 students and approximately £24 million for all UK providers that had suitable provider-level metrics for TEF in the 2017-18 cycle.

131. It has been clear that for some providers participating in the pilot, while absolute cost was low, a significant burden was placed on them and often responsibility for carrying out the varied demands of the pilot fell to a single person. This is not always an issue related to provider size; for example, some small specialist providers are well resourced. However, further education colleges that offer multiple subjects on a small scale and whose main business is not higher education can be identified as a group that will benefit from support in future, and particular attention should be paid to this group in the second subject pilot. The level of burden for these providers will interact with decisions relating to how to deal with subjects with limited data.

Size of future exercise

132. The pilot tested each model with 31 providers and in the time available it was sensible to concentrate on fundamental model design questions. However it is important to note that the full subject TEF will be much larger in scale, with substantial delivery implications.

133. We currently estimate that a full-scale exercise would require in the region of 3,500 to 4,500 subject assessments. These numbers will be impacted by the numbers of providers who successfully register with the OfS, as well as multiple decisions including:

- how to manage non-reportable data and small cohort sizes, while ensuring as broad coverage as possible
- the ongoing discussions with the Higher Education Statistics Agency (HESA) about refining the CAH2 architecture.

134. In the pilot, some panels had significantly higher workloads and this will need addressing at full scale. We estimate that larger panels may, at the top end, be required to produce in the region of 400 ratings. In the pilot assessment costs (covering venue costs and panel members and TEF officer fees and costs) were just below £1 million. Assessment costs would increase in a full exercise and it will be important to ensure that individual panel member workloads and commitments are reasonable.

135. In preparation for these increases in volume, the next pilot will need to test scalable assessment processes for:

- more formal and stringent data amendment and verification processes
- more thorough training and panel calibration exercises, and formal cross-panel moderation
- producing and quality assuring qualitative outcomes (a report similar or equivalent to a statement of findings).

Potential impacts

136. Two significant themes emerged from discussions with providers and panel members about the potential impacts of TEF at subject-level: its influence on provider behaviour, and communicating the meaning of subject ratings to applicants and current students.
137. There was a high level of consensus among participating providers that subject-level TEF will have a very positive impact on enhancement. A number of providers reported that access to subject-level metrics and the pilot exercise itself has already been used to drive enhancement and inform internal quality processes. But as the Main Panel Chair's report suggests, attention should be paid to potential impacts on widening participation and innovation, particularly in relation to new courses which lack data for assessment.
138. Subject-level ratings will help inform applicant choice by providing more granular information than the current provider-level TEF¹⁷. Ratings at subject-level will also create a new set of communications challenges. Both panel members and providers highlighted the need for careful and consistent handling of messaging to avoid any unintended consequences in the way applicants and current students interpret subject ratings. This was felt to be particularly important for subjects with 'No rating', and where the subject rating differs from the provider-level rating. It was recommended that the OfS could address the issue by carrying out further research and user testing in order to best understand how to present subject TEF information.

¹⁷ Research commissioned by the DfE and conducted by IFF Research indicates that applicants and current students would value the introduction of subject-level TEF ratings (see <https://www.gov.uk/government/publications/teaching-excellence-framework-and-informing-student-choice>, paragraph 1.37).

Summary ratings and analysis

This section provides summary information about the indicative ratings that were generated in the first subject pilot. The information has been anonymised, de-identified and aggregated in accordance with OfS data retention and publication policies.

General patterns

139. Summary charts are provided for:

- projected sector (all providers across the UK with available data):
 - subject-level initial hypotheses
- Model A:
 - provider-level outcomes
 - subject-level initial hypotheses, assessed subjects only (this includes subjects generated as exceptions and subjects that were not exceptions but were assessed to test the validity of model A)
 - subject-level outcomes for assessed subjects only
 - subject-level initial hypotheses for all subjects (this includes subjects generated as exceptions, subjects that were not exceptions but were assessed to test the validity of model A, and subjects that were not exceptions and inherited the provider rating)
 - subject-level outcomes for all subjects
- Model B:
 - subject-level initial hypotheses
 - provider-level outcomes
 - subject-level outcomes for all subjects.

140. These charts demonstrate the distribution of final ratings and the movement of these ratings from the initial hypothesis. Readers should exercise caution in interpreting these results given that one of the clear findings arising from the pilot is that neither model was fully fit for purpose for generating ratings. Additionally, analysis of ratings can provide only limited confidence in the predictability of patterns observed. It is not possible to infer that the distributions shown here would be replicated in a full scale assessment of the sector at large for the following reasons:

- a. Sample sizes within the pilot have been relatively small, especially when they are considered at the level of subject or assessment panel.

- b. The selection of providers to participate in the pilot oversampled some types of provider to capture those with distinctive features. This means that the pilot results should not be interpreted as representative of the whole sector.
- c. While the sample as a whole was selected to ensure a range of characteristics were represented, each panel had a unique sample in both models. Figure 15 (Model A) and Figure 23 (Model B) show how these samples differed in terms of initial hypotheses profiles but other factors varied, such as the propensity for non-reportable data and subject size. The profile of results and the movement of ratings should therefore be expected to differ between panels. The only results which may be directly compared are those relating to providers that took part in both models (see paragraphs 140 to 141 and Figures 26 and 27).
- d. While moderation was undertaken (through data analysis, observation and Main Panel discussion) it was light-touch and more formal and systematic moderation would be required at full scale.

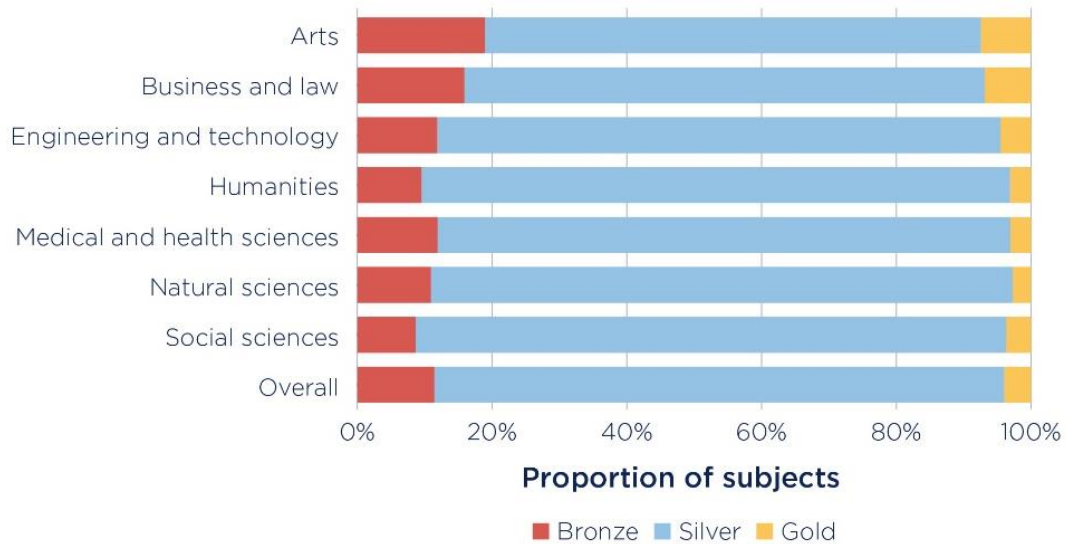
141. Nonetheless, the following patterns can be observed across the data:

- a. At provider-level, the distribution of Bronze, Silver and Gold ratings broadly follows that observed in the TEF Year Two and TEF Year Three exercises.
- b. There are differences in this distribution at subject-level. Firstly, it is clear that larger proportions of Silver ratings are observed across all subjects than at provider-level. Secondly, there are differences in distributions between subjects. Some of these differences may be related to the features of the pilot participants, but it should be noted that a propensity towards Silver is also present in the profile of initial hypotheses across all subjects.
- c. All panels see outcomes that move from the initial hypothesis, demonstrating the importance of written evidence and panel judgement.

142. The OfS undertook an analysis of these results including, for both Model A and Model B, employing binary logistic regression models to attempt to identify relationships between student characteristics at the provider (using contextual data) and the ratings of subjects assessed. The models had limited capacity to identify effects given the sample sizes involved when observing outcomes for subjects within individual providers (rather than across providers). They did provide some suggestion that the proportions of BME students, mature students and of students with low or no entry qualifications impacted on the propensity of a subject to gain Bronze or Gold ratings under Model A. However, there is no consistent overall pattern and it would be inappropriate to assume that the results would be replicated across a wider range of providers.

Projected sector: Subject-level initial hypotheses

Figure 12: Profile of step 1a initial hypothesis at subject-level, by subject panel (whole sector)



Model A provider-level outcomes

Figure 13: Model A – profile of final provider-level ratings

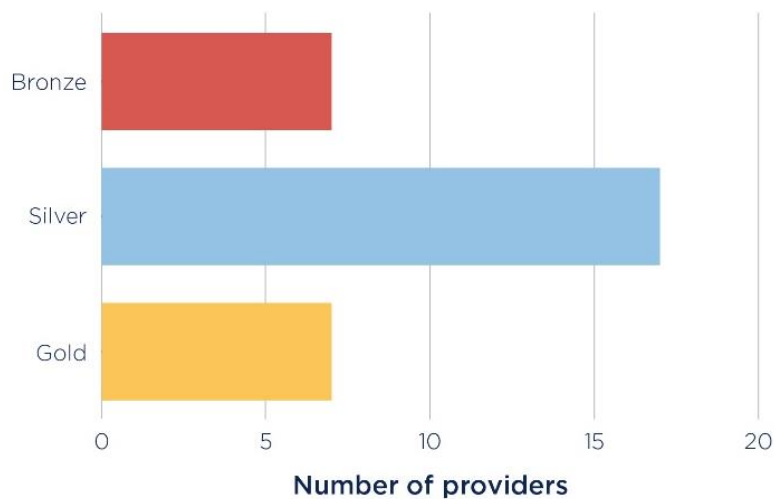
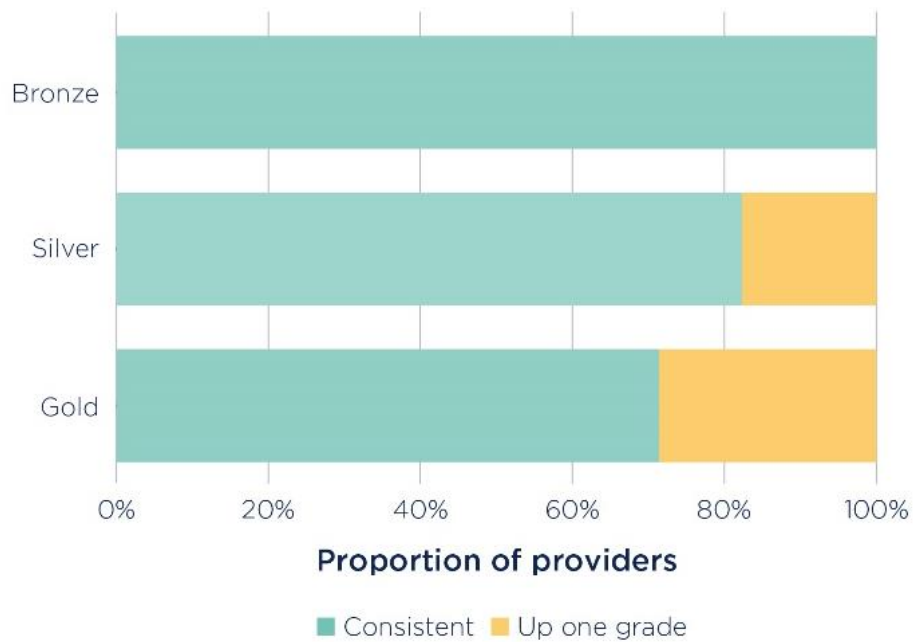


Figure 14: Model A – difference between providers’ step 1a initial hypothesis and final rating



Model A subject-level initial hypotheses and outcomes – assessed subjects only

Figure 15: Model A – profile of step 1a initial hypothesis at subject-level (subjects that were panel assessed only)

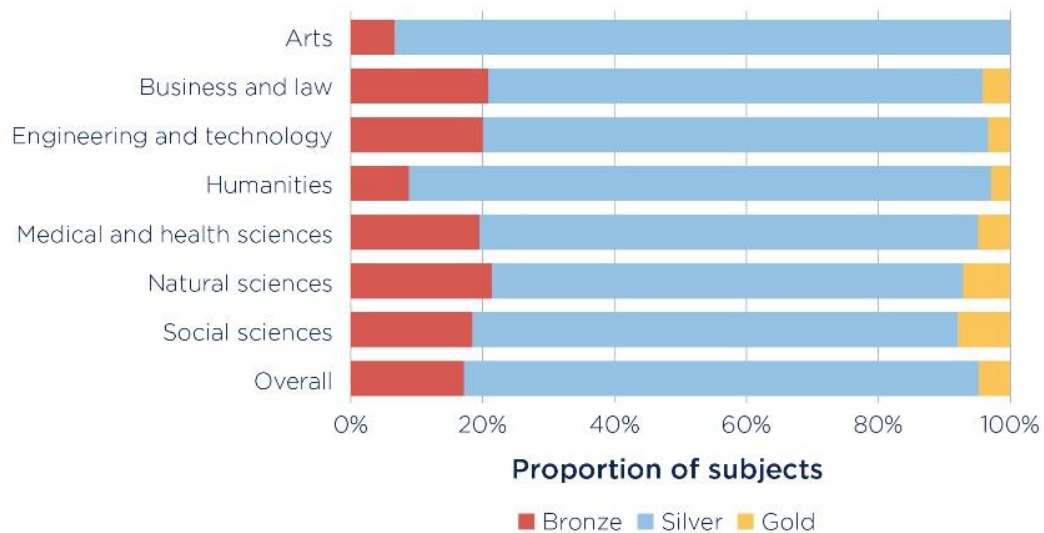


Figure 16: Model A – profile of final ratings at subject-level (subjects that were panel assessed only)

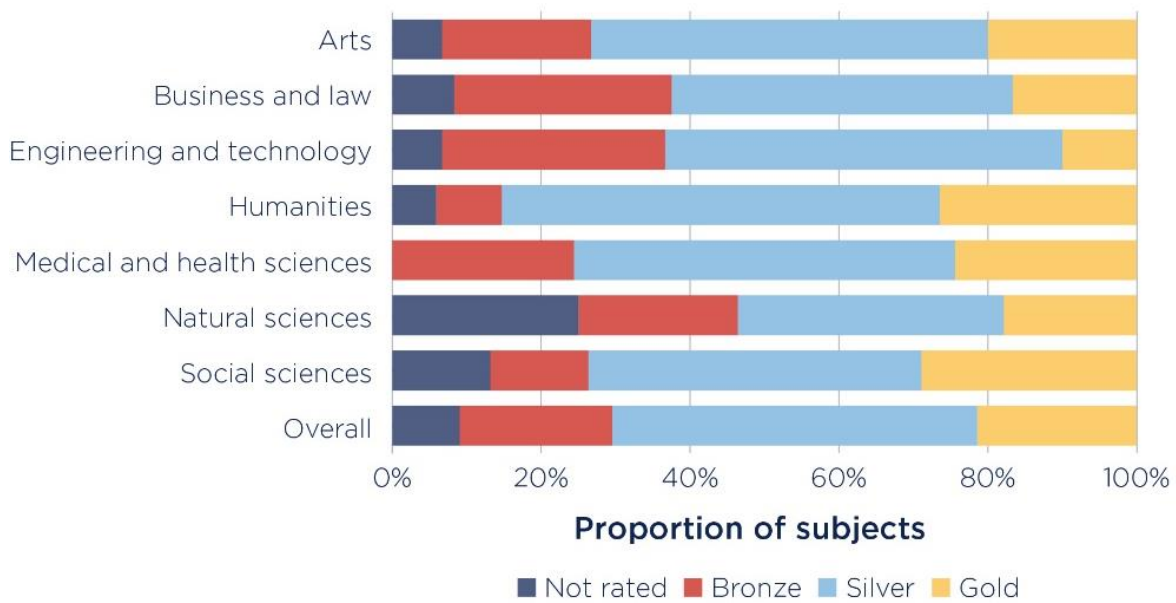
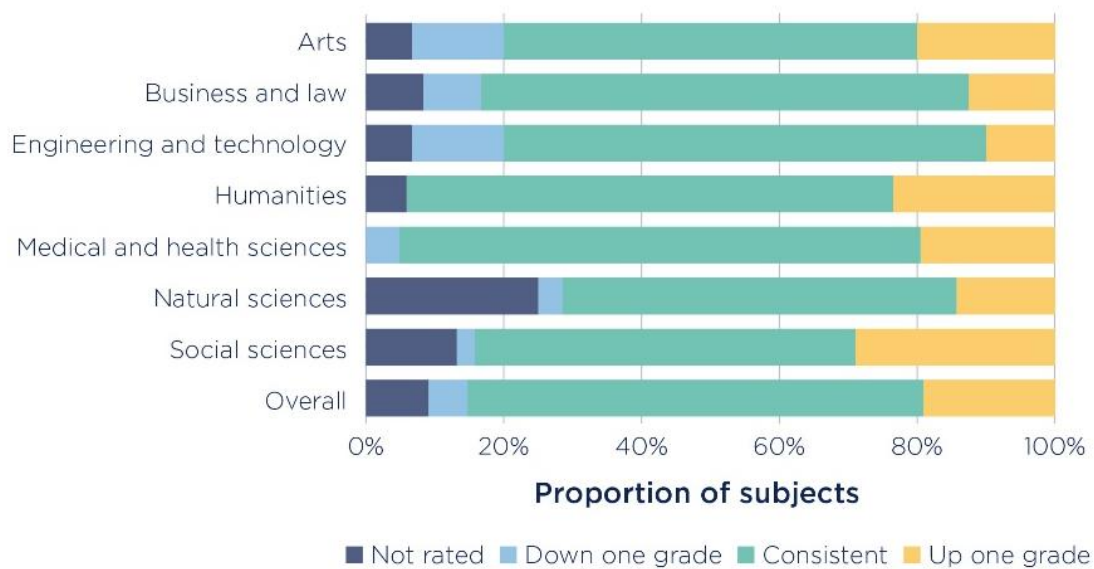


Figure 17: Model A – Difference between subject’s step 1a initial hypothesis and final rating (subjects that were panel assessed only)



Model A subject-level initial hypotheses and outcomes – all subjects

Figure 18: Model A – profile of step 1a initial hypothesis at subject-level (all subjects)

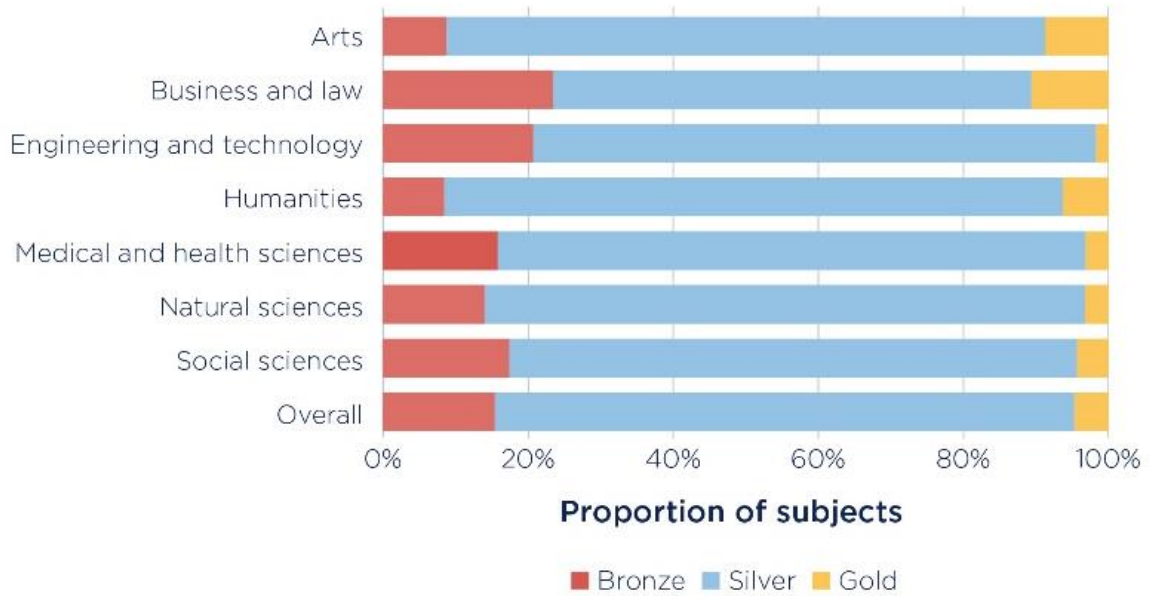


Figure 19: Model A – profile of final ratings at subject-level (all subjects)

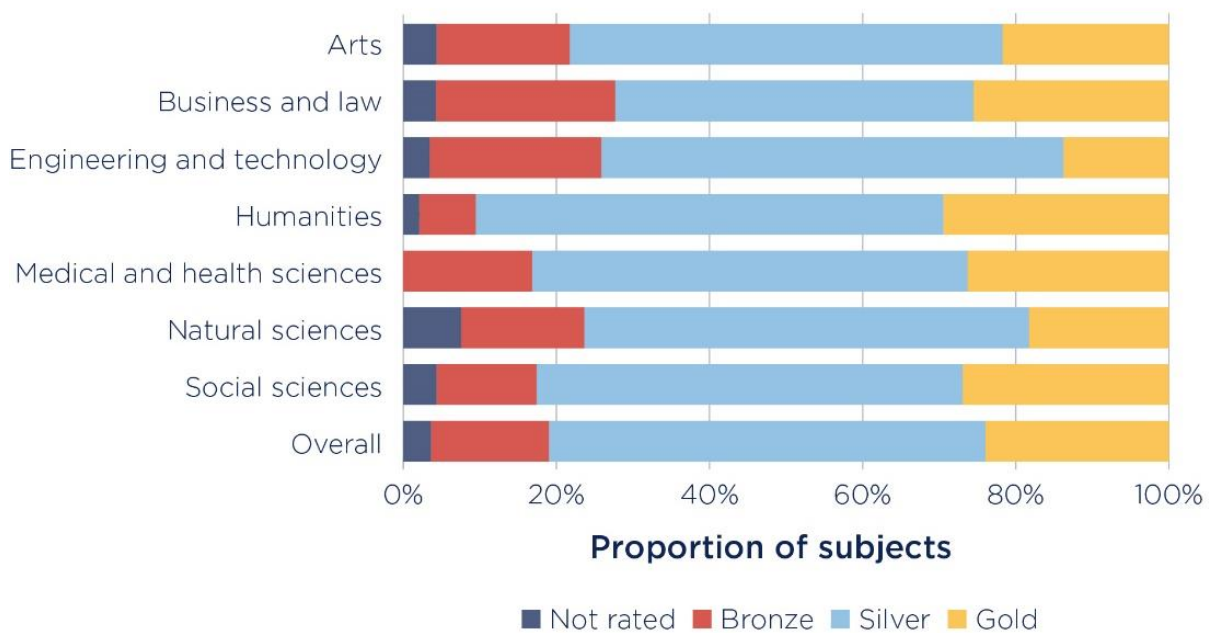
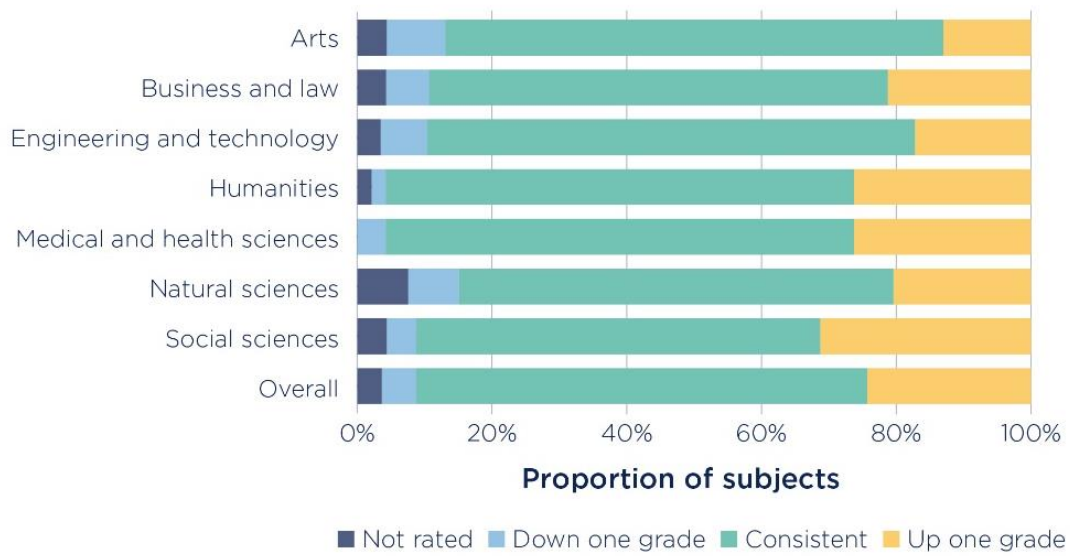


Figure 20: Model A – Difference between subject’s step 1a initial hypothesis and final rating (all subjects)



Model B provider-level outcomes

Figure 21: Model B – Profile of final provider-level ratings

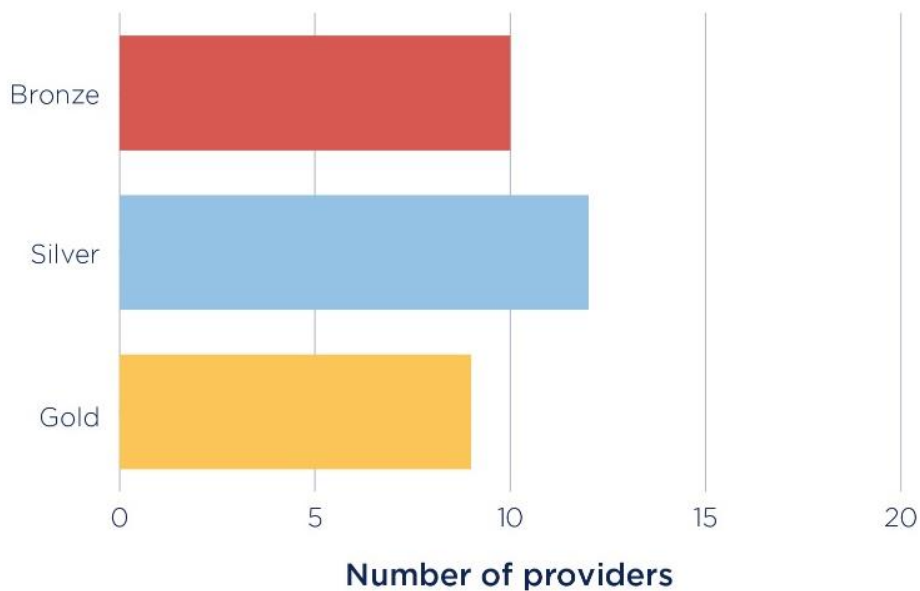
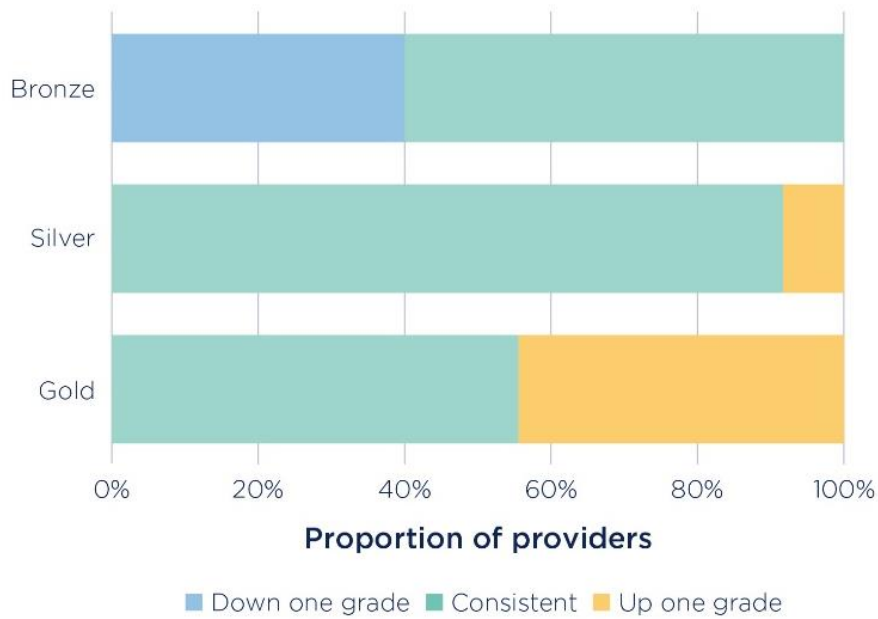


Figure 22: Model B – Difference between provider’s step 1a initial hypothesis and final rating



Model B subject-level outcomes – all subjects

Figure 23: Model B – Profile of step 1a initial hypothesis at subject-level

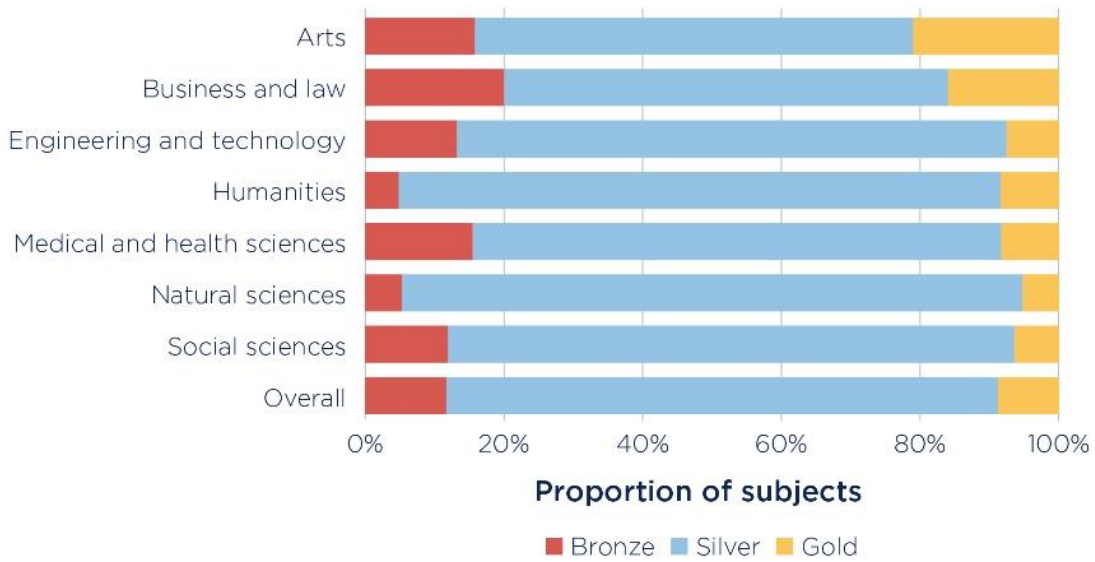


Figure 24: Model B – Profile of final ratings at subject-level

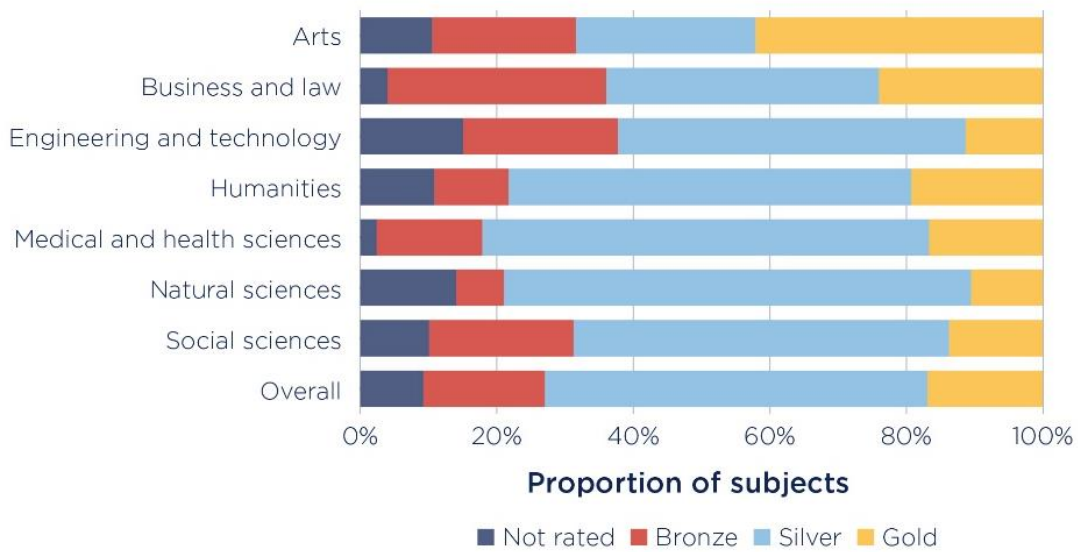
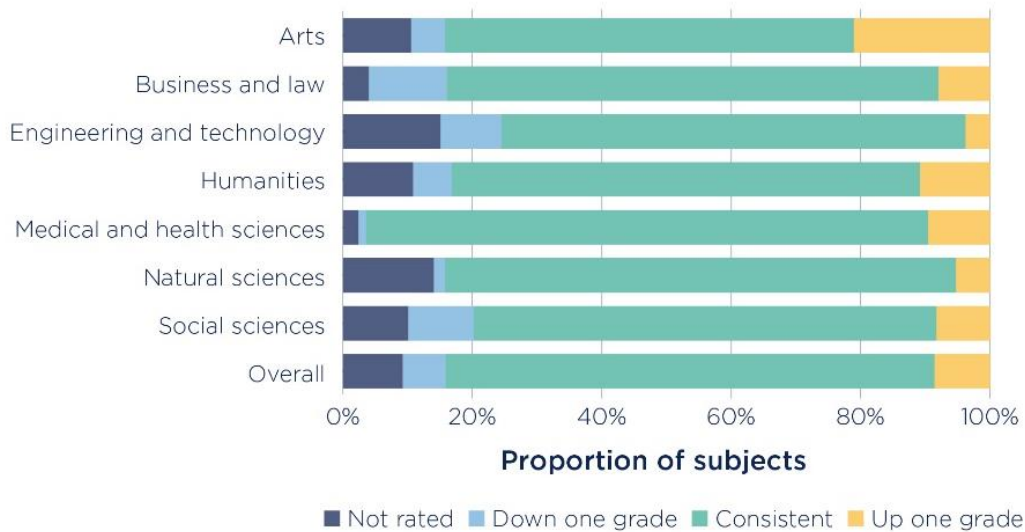


Figure 25: Difference between subject’s step 1a initial hypothesis and final rating



Results for providers in both models

143. A key feature of the pilot design was the inclusion of 12 providers that participated in both models, which allows for direct comparison of the ratings generated. Figures 27 and 28 illustrate the distribution of ratings for these providers between the models, and Table 6 shows the actual differences per subject. It compares the rating awarded under Model B with that under Model A ((that were panel assessed only).

144. OfS analysts examined whether flag profiles had an impact on whether subjects were rated differently across the models, but found no effect. The more likely conclusion, therefore, is that

differences arise as a result of the different subject-level submission formats between the models and additionally the subject-based initial hypothesis at provider-level in Model B.

Table 6: Proportion of positive, negative, and neutral changes from Model A to Model B (subjects that were panel assessed only)

Level	Rating difference	Ratings count	Total number of ratings
Provider	Lower rating in Model B	2	12
Provider	Equal rating in both models	10	12
Subject	Lower rating in Model B	6	70
Subject	Equal rating in both models	60	70
Subject	Higher rating in Model B	4	70

Final provider and subject-level ratings for providers in both models

Figure 26: Profile of provider-level ratings by model (providers in both models only)

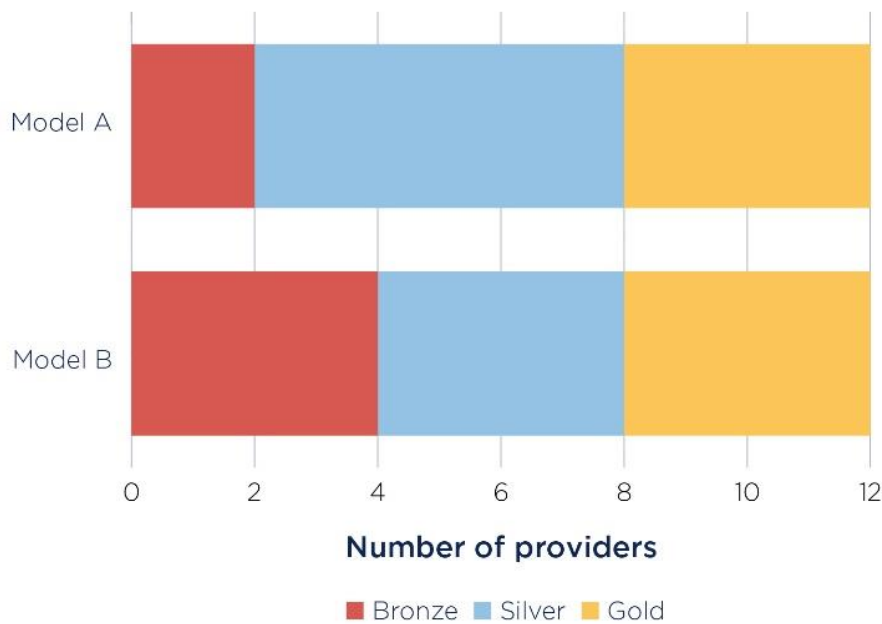
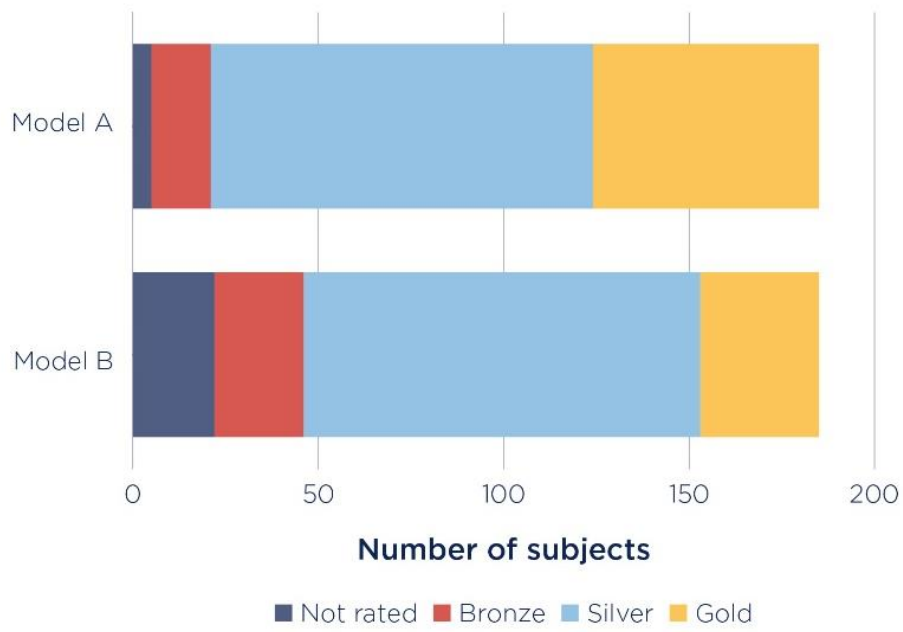


Figure 27: Profile of subject-level ratings by model (provider in both models only), (all subjects)





© The Office for Students copyright 2018

This publication is available under the Open Government Licence 3.0.

www.nationalarchives.gov.uk/doc/open-government-licence/version/3/