# Teaching Excellence and Student Outcomes Framework (TEF)

## Findings from the subject-level pilot 2018-19

**Enquiries to:** tef@officeforstudents.org.uk

**Publication date** 21 January 2021

# Foreword

**By Professor Janice Kay, CBE**
**TEF subject-level pilot chair**

From the start, the vision for TEF has been that it would quickly become a vital component of our higher education landscape and that it would drive better outcomes for students. It is clear that TEF has led to universities and colleges increasingly taking a more systematic and effective approach to enhancing teaching and learning, which in turn will lead to better experiences for students. This growing focus will benefit both current and future students, as well as the wider UK higher education environment. We should build on these strong foundations.

It is always exciting to be involved from the beginning in something of such importance. So, it has been a privilege to chair the TEF subject-level pilots, alongside many dedicated colleagues and students who have worked tirelessly on the project. The pilots sought out ways of providing even better information to support student choice. It has also shown how we might incentivise continuous improvement and enhancement across the sector. We took full advantage of this pilot phase, scrutinising, evaluating and testing a large variety of different mechanisms that would be required to operate an effective large-scale subject-level TEF.

What did we learn? The model itself is important, and this year's approach of assessing all subjects was more effective than the two previously tested. On a relatively small scale, we were able to run a successful subject-level pilot with a small number of higher education providers. We saw how variability of subject provision exists within institutions, as well as across them. It is clear that there is value in scrutinising subject provision as a vital part of overall assessment of the quality, strategy and management of universities and colleges.

The pilots surfaced challenges to operating a full subject-level exercise that require further development of the TEF framework. Important themes emerged around the complexity of data and the importance of student engagement. How do we deal with the reality of technical statistical issues? The chair's report[1] shows how we worked systematically through many of the questions that arose. How can and should student evidence be factored into our decisions? Through the outstanding work of students on our panel, and in pilot providers, we made real progress in working in partnership. There are no simple solutions, but this report thoroughly details the challenges, how we responded to them and where further development is needed.

The overall message that should be taken from the exercise is one of optimism: in drilling down to subject-level information we were presented with a snapshot of the sector that was bursting with

[1] The report from the TEF main panel chair is available as Annex A at www.officeforstudents.org.uk/publications/tef-findings-from-the-second-subject-level-pilot-2018-19/.

examples of institutional excellence. The exercise was also constructive in focusing attention on areas of provision that could be improved.

Given the remaining challenges, it is important that a careful plan for developing the next exercise is put in place. But it will be worth the time spent now as the reward will be a meaningful and mature exercise, fit for the future.

The learning from the pilots has resulted in some excellent, thorough evaluation of where TEF stands. Alongside the outcomes of Dame Shirley Pearce's review, I know that the exercise can be taken forward in a very positive way. I encourage colleagues to continue to be bold and aspirational in their vision for what a truly excellent higher education sector looks like, one that is the envy of the world.

# Contents

# Executive summary

1. During academic years 2017-18 and 2018-19 the Office for Students (OfS) piloted models for producing Teaching Excellence and Student Outcomes Framework (TEF) ratings at subject level.

2. The pilots built on the successful delivery of the TEF exercises that have produced publicly available ratings for higher education providers (referred to as 'provider-level TEF') since 2017. Through the pilots, we tested how to produce ratings for each undergraduate taught subject[2] at a university or college (referred to as 'subject-level TEF'). To deliver the pilot we recruited an independent panel and a range of providers to test the proposals and were supported during the assessment phase by colleagues from the Quality Assurance Agency for Higher Education (QAA).

3. This report collates the results and key findings from the 2018-19 TEF subject-level pilot, which was the second year of piloting delivered by the OfS. It is an OfS report informed by a series of supporting independent reports as well as OfS analysis conducted following the conclusion of the pilot. These are listed on page 9 and available on the OfS website.[3]

4. At the same time as the second pilot was conducted, the Secretary of State for Education commissioned the statutory independent review[4] of the TEF (in November 2018). This findings report focuses on what we have learned from testing the pilot model. It does not seek to address wider questions about the TEF that were within the remit of the independent review, or propose solutions that would pre-empt the review's recommendations.

5. The findings show significant variation in the experience and outcomes of students studying different subjects, within individual universities and colleges. This variation would remain hidden if assessments only consider provider-level evidence. Subject assessment has highlighted individual subjects where the excellent experience and outcomes their students receive is beyond that apparent across the provider as a whole, as well as individual subjects where there is room for improvement. Understanding and addressing this variation within universities and colleges has the potential to drive excellence in learning and teaching, and to provide more useful information to students.

6. The model for the second pilot was an improvement on previous models, and significant progress was made in how assessment could be extended to subject level. The universities and colleges involved in the pilot engaged constructively to help us to explore issues and develop solutions, and changes to the process meant students were able to be more actively involved than before. The panels developed substantial expertise and the processes relating

[2] The subjects assessed in the subject-level pilot were based on the HESA Common Aggregation Hierarchy at level 2 (CAH2). In this pilot, an amended version of the CAH2 was used, based on feedback from the previous year of pilots. Annex A of the 'TEF subject-level pilot guide', available at www.officeforstudents.org.uk/advice-and-guidance/teaching/subject-level-pilot-2018-19/, sets out the version of CAH2 used in this pilot. Further information on the development of the new HECoS and the CAH is available at www.hesa.ac.uk/innovation/hecos.

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

[4] See: https://www.gov.uk/government/publications/independent-review-of-tef-report.

to preparation, training and assessment were refined. Our key findings, set out below, illustrate the many areas where we have learned valuable lessons.

7. Whilst there is scope to extend the evidence and assessments within TEF to subject level, the pilots and further metrics analysis show that we do not yet have a model for producing robust subject-level ratings across the sector that is ready to implement. There are significant limitations in the data at subject level, and other issues that would require further development before moving to a scheme that systematically rates all subjects.

8. The OfS will carefully consider the findings from this report, alongside the recommendations from the independent review and any guidance provided by the Secretary of State, and we will consult on a proposed approach in 2021.

There are no immediate actions arising from this report for higher education providers. Providers, students and other stakeholders will be invited to respond to the forthcoming consultation on the future rating scheme later in 2021. Further action will be required once information about the next iteration of the rating scheme is published by the OfS.

## Key findings

9. For simplicity, higher education providers including universities and colleges are referred to as ‘providers’ throughout this report.

10. Detailed commentary on these key findings is available from page 24 onwards.

- **The comprehensive model piloted in 2018-19 was an improvement on the previous year’s pilot models.** The revised model was comprehensive in that it ensured that all subject ratings were based on a full assessment of that subject. This addressed design issues with previous models, where some ratings were made with limited evidence or no direct assessment. However, the precise relationship between subject-level and provider-level ratings would need further consideration.

- **The pilot revealed significant variation in performance within providers, confirming the importance of taking account of this variation in the TEF outcomes.** It was common for individual providers to receive the full spectrum of ratings (Gold, Silver and Bronze) across their subjects. Scrutiny at subject level identified areas of excellence and areas for improvement within providers, in a way that assessing only provider-level evidence does not.

- **There was a clear expectation that subject-level TEF would help drive improvements in higher education.** Participating providers, students and panels reported that analysing performance at subject level would have positive benefits and incentivise excellence.

- **The TEF needs to be better aligned with other OfS functions.** Participants reflected that the TEF would have more internal impact if processes were more closely linked to other regulatory functions (in particular access and participation plans and ongoing monitoring against the other quality and standards conditions of registration). This would also enable providers to respond more efficiently. Better alignment would also help clarify how to interpret information that is relevant across these different functions.

- **Participants engaged constructively.** As they worked through the various processes, participants helped us to identify and explore a number of challenges involved in subject-level assessment. They engaged constructively in seeking to work through them, and whilst some issues remained unresolved, a great number of their ideas and suggestions helped to solve problems and refine processes.

- **Improvements were made to student engagement in the process and there was desire to extend this further in future.** The increased focus on student voice and student partnership in the assessment was welcomed. We trialled ways of increasing student involvement in TEF submissions and some progress was made particularly at provider level. However, there were practical barriers such as time and resource constraints, and involvement at subject level was patchy. Across panels and student representatives there was a desire to address these barriers, and further extend the opportunities for students to submit evidence more directly into the process.

- **Pilot providers and panels encountered significant limitations in the data at subject level.** These included:
  − metrics that were missing or unreportable
  − data that was based on small cohorts and with limited statistical reliability
  − data that no longer related to the courses being offered in a subject.

  We tested means of mitigating these limitations but found that they did not resolve the issues. The panels needed a disproportionate amount of time deliberating over cases with limited evidence; they made judgements in each case on whether there was sufficient evidence to determine a rating, but this process was inconsistent and would not be scalable.

- **Further analysis shows that many subjects across the sector do not have large enough student cohorts for the current metrics flagging method to robustly inform assessments.** Analysis shows that under the current method for generating flags and an initial hypothesis, a subject would need to cover several hundred students in order for the metrics flags to robustly inform the assessments. Many subjects across the sector have smaller cohorts than this. While panels had a general awareness that metrics based on smaller cohorts had limitations, they conducted their assessments without knowledge of how large the cohorts needed to be to consistently generate flags, as this analysis was only carried out after the assessments had concluded. This could be addressed, though, through further work to improve how metrics could be generated and used robustly at subject level.

- **Improvements to the subject categories following the previous year's pilot were welcomed, and some further refinement would be needed in this area.** Although improvements were made to the subject categories, some mismatches with internal provider structures remained. Where provision did not map to the subject categories, providers experienced difficulties in interpreting their metrics and writing submissions, and the ratings they received were of limited value for enhancement. There were also concerns that fixed subject categories would constrain innovation. While some providers suggested they should have flexibility in categorising their subjects, no solutions have been proposed at this stage that would be scalable. Further work would be needed to assess the scale of the mismatches and to identify further improvements.

- **Concerns remained about the treatment of interdisciplinary provision.** Only one-third of subjects assessed in the pilot were completely self-contained (i.e. with all students mapping to a single subject category). The majority included students studying in multiple subject categories, and around 15 per cent of subjects included students also studying in five or more other subject categories. We trialled the use of additional data and expertise within panels which helped to take account of interdisciplinary provision. While this was found to be helpful, challenges remained.

- **The exercise was complex, and we identified some data that could be omitted in future.** We piloted additional metrics and contextual data, most of which were welcomed. However, the overall package of information became unwieldy and the metrics involve inherent complexity, requiring users of the data to be trained and to develop specialist TEF expertise. This made it challenging for providers to involve people across their departments, and challenging for smaller providers with limited capacity to develop this

expertise. Tailored guidance for different roles within providers was suggested. Panel members faced a steep learning curve. They found that some of the data was of little use and could be omitted. In particular, the employment maps and contextual data could be omitted if regional employment factors were to be taken into account within the metrics.

- **The metrics relating to employment outcomes and differential attainment need further development.** There was a broad view that, if used, LEO metrics should in future take account of regional factors. The pilot tested how experimental 'supplementary' data indicating gaps in degree attainment between groups of students, could be used in provider-level assessment. The main panel agreed that differential attainment could be considered within the TEF, as a key issue to be addressed in the sector. However, the panel found that to make use of the data in assessment, it would need to be developed further into a metric. Also, further clarity would be needed on how it relates to our regulation of access and participation.

- **The exercise demonstrated that more time would be needed throughout the process.** The complexity of the exercise was exacerbated by the tight timescales. This hampered the ability of providers to brief staff, analyse data, write and coordinate submissions across departments, and involve students. The panels achieved broad consistency in their decisions, but were unable on the pilot timetable to fully moderate their judgements. More time would be needed in a full exercise for all participants: for providers to write submissions, for students to engage in the process, for panels to be trained and conduct their assessments, for calibration, moderation and the consistent production of statements of findings, and for the OfS to develop systems, process data and deliver the exercise at scale.

- **Care would need to be taken to ensure subject-level assessment is fair for further education colleges (FECs) and smaller or specialist providers.** There were apparent differences in the organisational capacity and ability of different types of providers to engage in subject-level data and submissions. The pilot outcomes also differed by type of provider. These issues would need to be explored further to ensure subject-level assessment is as fair as possible to all participants.

11. Throughout this report readers should exercise caution in interpreting results given the statistical issues raised and that the providers participating in the pilot were selected for their diversity, and so may not be representative of the final set of providers taking part in the future exercise.

## Supporting documents

**Supporting documents are available alongside this document on the OfS website:**[5]

**Annexes: Panel reports**

**Annex A:** Main panel report

**Annex B:** Student findings report

**Annex C:** Subject panel reports

**Annex D:** Widening participation expert report

**Annex E:** Employment expert report

**Annexes: OfS analysis and glossary**

**Annex F:** TEF subject-level pilot outcomes analysis

**Annex G:** Logistic regression models

**Annex H:** Analysis of Type I and II errors in TEF flags

**Annex I:** TEF subject-level pilot glossary

**Independent report by IFF Research**

TEF subject-level pilot evaluation – Provider perspectives

[5] These reports are available at: www.officeforstudents.org.uk/publications/tef-findings-from-the-second-subject-level-pilot-2018-19/.

# Introduction

## Policy context

12. The TEF was introduced by the Department for Education (DfE) in 2016 to recognise and reward excellent teaching in UK higher education and to inform student choice.

13. The current TEF assesses undergraduate provision at 'provider level'; a single rating is awarded to each higher education provider that takes part. The May 2016 White Paper 'Success as a knowledge economy'[6] set out the Government's intention for TEF assessments in future to be carried out at subject level.

14. The initial provider-level TEF exercises and a first subject-level pilot were delivered by the Higher Education Funding Council for England (HEFCE) and then, from April 2018, by the OfS, according to the DfE's specifications. (For simplicity, we will refer to the OfS in place of HEFCE throughout this report.)

15. The OfS was established by the Higher Education and Research Act 2017 (HERA)[7] and became fully operational in April 2018. The OfS has adopted TEF under HERA section 25(1) as its rating scheme, to support its overall strategic aim to ensure that every student, whatever their background, has a fulfilling experience of higher education that enriches their lives and careers.

16. During 2018-19, the OfS ran a second pilot of a subject-level TEF, building on the lessons from the first pilot to develop and test a potential model for producing TEF ratings at subject level as well as provider level.

17. At the same time as the second pilot was conducted, the statutory independent review of the TEF (commissioned in November 2018 by the Secretary of State for Education) was completed by Dame Shirley Pearce with the support of an expert advisory group. The review, as set out in section 26 of HERA, reported on the operation of TEF. The full findings of this review were published in the Independent Review of the Teaching Excellence and Student Outcomes Framework (TEF) Report to the Secretary of State for Education[8] released in January 2021.

18. The OfS will continue to develop and implement a rating scheme, informed by the independent review, the Government's response to the independent review, and the findings of the second pilot set out in this report.

## Acknowledgements

19. We are grateful to the hundreds of colleagues from across the sector who have shown dedication and commitment in delivering the pilots, working collaboratively with us to explore,

[6] See: www.gov.uk/government/publications/higher-education-success-as-a-knowledge-economy-white-paper.

[7] See: https://www.legislation.gov.uk/ukpga/2017/29/contents/enacted.

[8] Independent review of the Teaching Excellence and Student Outcomes Framework (TEF). See: https://www.gov.uk/government/publications/independent-review-of-tef-report.

problem solve and troubleshoot issues across both years of the pilot. We would like to thank the 82 providers, 175 panel members, and colleagues from the QAA for their involvement, and to highlight the unique and valuable contribution made by the student representatives based at pilot providers, and the 49 student panellists.

20. Our particular thanks go to the main panel chair, Professor Janice Kay, deputy chairs, Helen Higson and Josh Gulrajani, and all of the subject panel chairs and deputies for their leadership, insight and advice.

## Aims and scope of the second pilot

21. The first year of subject-level TEF piloting took place in 2017-18. It tested two models of subject and provider-level assessment and followed the DfE's TEF subject-level pilot specification[9], published in July 2017. Each of the two models were designed to limit the burden involved in subject-level assessment in a different way. However, the first pilot showed that neither model produced sufficiently robust ratings at subject level, and that a more comprehensive approach should be tested. A full evaluation report 'Findings of the first subject pilot, 2017-18'[10] was published in October 2018.

22. Building on these findings, the 2018-19 pilot tested a single model in which all subjects at the participating providers were assessed in full. The second pilot also trialled potential refinements to the criteria, metrics and other processes that were not explored in the 2017-18 pilot, including the enhancement of student engagement with the submission process and production of 'statements of findings' at subject level. We also considered more fully the implications for scaling up subject-level assessments to the whole sector.

23. In October 2018 we published the 'TEF Subject-level pilot guide'[11] which sets out in full the aims and methods used in the pilot. This evaluation report briefly summarises the pilot method (in the 'how the second pilot was conducted' section); for further information please refer to the Subject-level pilot guide.

24. In the second pilot, a diverse group of 45 providers took part. Panels comprising 147 members assessed 45 provider-level submissions and 630 subject-level submissions. A full-scale subject-level exercise (if using the same approach as the second pilot) would involve up to 3,500 subjects across approximately 400 providers.[12]

25. Both pilots were developmental exercises: our focus has been on the application of the models and methods of assessment, not on the performance of individual providers that took part. This is reflected in the findings in this report. Whilst participating providers each received

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

[10] TEF: Findings from the first subject pilot, 2017-18. Available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-findings-from-the-first-subject-pilot-2017-18/.

[11] TEF subject-level pilot 2018-19 guide. Available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-subject-level-pilot-guide/.

[12] These estimates represent an upper limit based on the number of providers and subjects that would currently be eligible for assessment and which meet the minimum data thresholds used in the second subject-level pilot.

their own indicative ratings, no ratings are published in a manner that could identify an individual provider.

## How the pilot was conducted

26. The TEF subject-level pilot guide[13] sets out the methods and procedures we followed. In summary, there were three phases of activity, with evaluation activities conducted throughout:

    a. **Set-up – September 2018 to October 2018.** During this phase we: recruited a sample of providers that reflected the diversity of the UK higher education sector; recruited the pilot panels; prepared guidance and training materials; and generated the pilot metrics, making some improvements to subject classifications.

    b. **Submissions – October 2018 to February 2019.** Providers received their metrics and prepared their submissions, with support and guidance provided by the OfS through a series of briefing events and supporting materials. Providers were expected to involve their students, and the OfS identified and provided briefings for a lead student representative at each provider.

    c. **Panel assessment – March 2019 to May 2019.** Having been trained in parallel with the submission phase, panel members assessed all submissions both individually and then collectively as panels, across a series of residential meetings. The main panel assessed provider-level submissions and the subject panels assessed subject-level submissions against the relevant criteria. The main panel carried out cross-panel consistency checks. Both levels of assessment produced outcomes of ratings and statements of findings. Providers were then given access to their outcomes in July 2019, to help inform evaluation.

## The pilot model

27. Each subject at the participating providers was assessed in full and, in parallel, so was the university or college as a whole. Many features of the existing provider-level TEF framework were used at both provider and subject level in the pilot:

    a. Independent peer review involving academics, students and other experts

    b. The assessment considered teaching excellence across three 'aspects of quality': teaching quality (TQ); learning environment (LE); and student outcomes and learning gain (SO), which are broken down into common criteria.[14] For example, 'Student engagement with learning' is a criterion within the teaching quality aspect.

    c. Holistic judgements based on the totality of the quantitative and qualitative evidence across all three aspects, set within the context of the provider and its students

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

[14] See the TEF subject-level pilot guide, available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-subject-level-pilot-guide/, Table 3 TEF criteria at provider and subject level, pages 24–25.

d. Each submission (whether at provider level or subject level) was rated on a three-point scale: Gold, Silver or Bronze

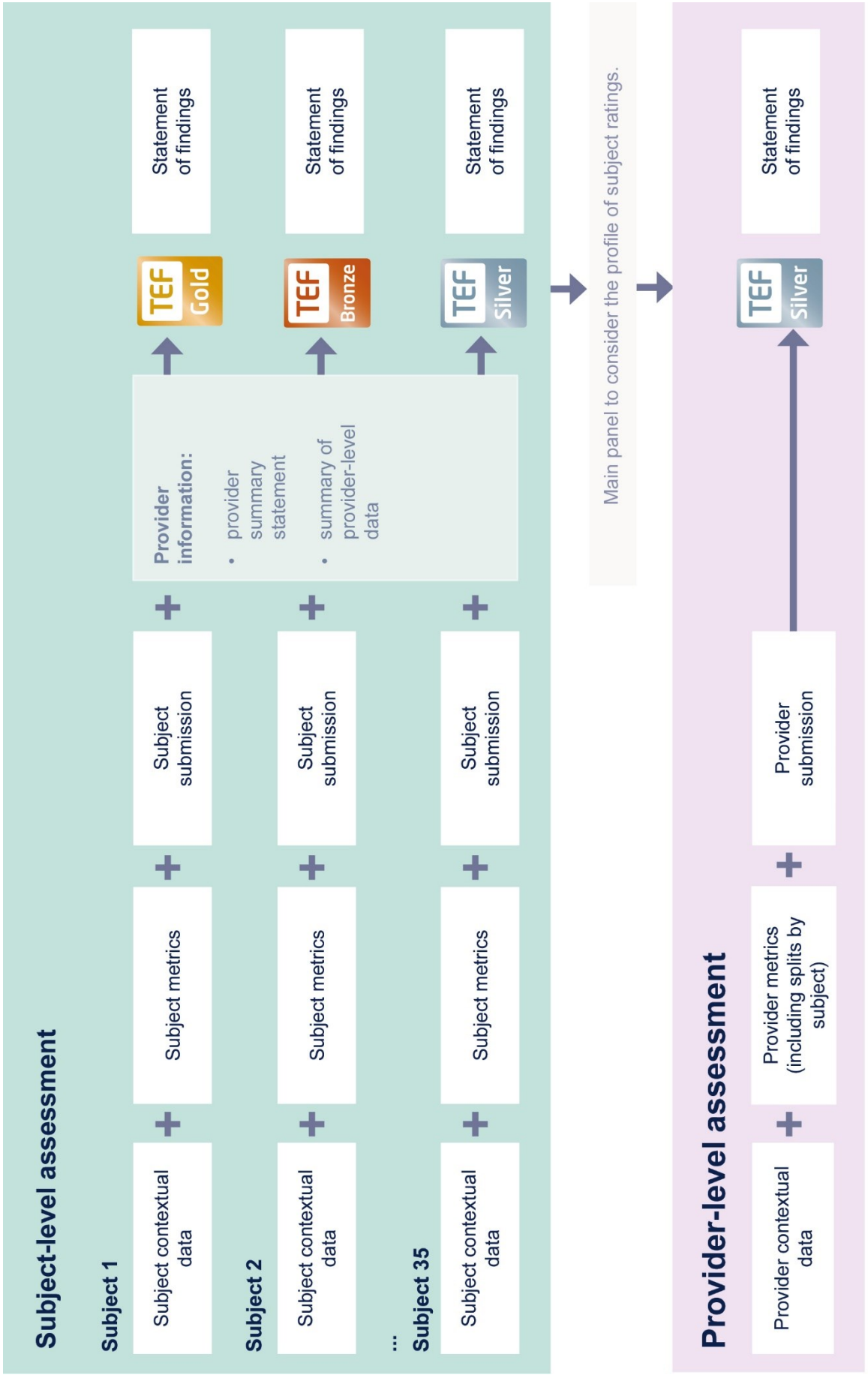
e. Students were involved throughout the process of submissions and assessment.

28. For the pilot, we tested the relationship between provider-level and subject-level assessments, including how subject-level information could inform the provider-level assessment. We also tailored processes, such as the criteria used to assess submissions, to recognise that there are different expectations and responsibilities at both provider and subject level.

29. In this pilot we introduced a student declaration, which was an opportunity for students to indicate how they were involved in their provider's submissions. These declarations were submitted directly to the OfS and were considered by the pilot main panel as additional contextual information.

30. The overall pilot assessment model is shown in Figure 1. Table 1 summarises the criteria, associated evidence and approach to assessment at both provider and subject level. Figure 2 summarises the process of submissions and assessment.

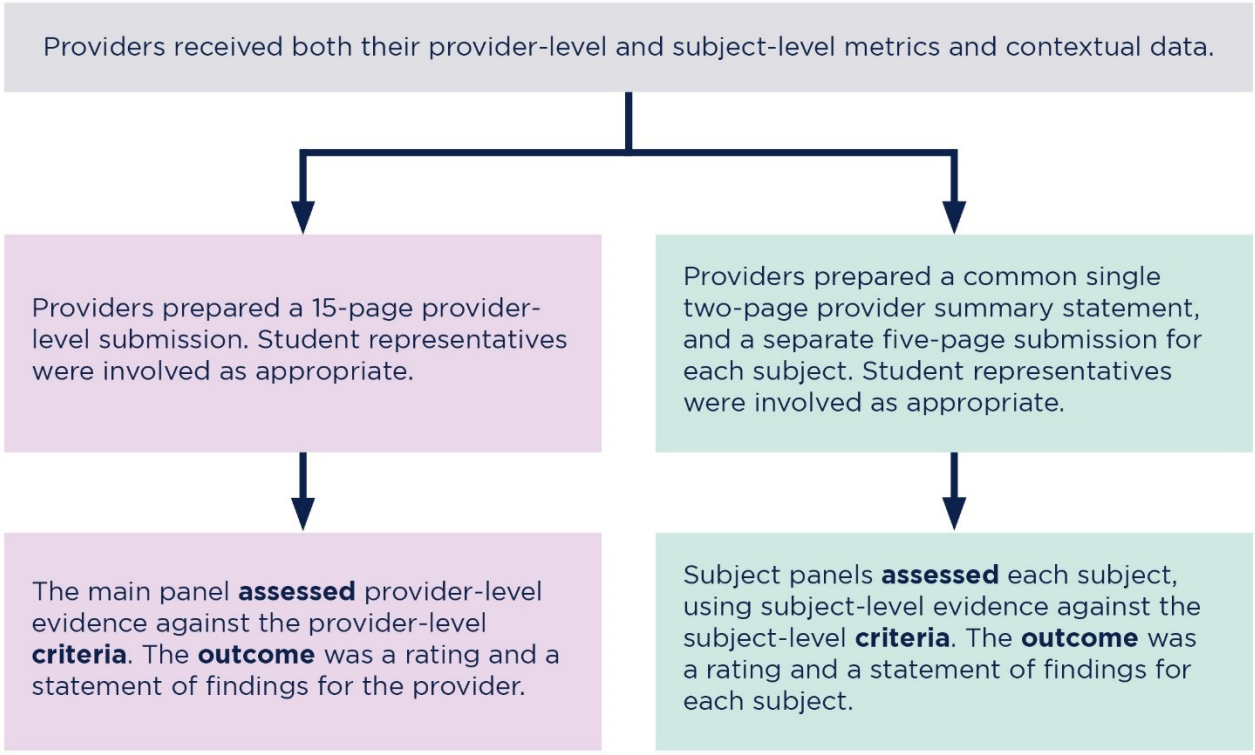
**Figure 1: Assessment model for the second pilot**



## Table 1: Pilot assessment framework

<table>
<tr><th colspan="2"></th><th>Provider level</th><th>Subject level</th></tr>
<tr><td colspan="2" rowspan="2">Criteria</td><td>Criteria defined at provider level:</td><td>Criteria defined at subject level:</td></tr>
<tr><td colspan="2">Teaching quality (TQ):<br>• Student engagement with learning (TQ1)<br>• Valuing teaching (TQ2)<br>• Rigour and stretch (TQ3)<br>• Feedback (TQ4)<br>• Student partnership (TQ5)<br>Learning environment (LE):<br>• Resources (LE1)<br>• Scholarship, research and professional practice (LE2)<br>• Personalised learning (LE3)<br>Student outcomes and learning gain (SO):<br>• Employability and transferable skills (SO1)<br>• Employment and further study (SO2)<br>• Positive outcomes for all (SO3)</td></tr>
<tr><td rowspan="4">Evidence</td><td>Contextual information</td><td>• Provider-level contextual data<br>• Maps of provider-level geographic context</td><td>• Subject-level contextual data<br>• Maps of subject-level geographic context</td></tr>
<tr><td rowspan="2">Metrics</td><td>Provider-level metrics:</td><td>Subject-level metrics:</td></tr>
<tr><td colspan="2">• Teaching on my course (NSS)<br>• Assessment and feedback (NSS)<br>• Student voice (NSS)<br>• Academic support (NSS)<br>• Learning resources (NSS)<br>• Continuation (HESA/ILR)<br>• Highly-skilled employment or higher study (DHLE)<br>• Sustained employment or further study (LEO)<br>• Above median earnings threshold or higher study (LEO)</td></tr>
<tr><td>Supplementary data</td><td>Where applicable:<br>• Differential degree attainment data and grade inflation data<br>• Additional data on part-time provision</td><td></td></tr>
<tr><td>Submissions</td><td>• Provider-level submission (up to 15 pages)</td><td>• Subject-level submission (up to five pages)<br>• Provider summary statement (up to two pages)</td></tr>
<tr><td colspan="2">Assessment</td><td>Assessment by the pilot main panel:<br>• Three-step method of assessment<br>• Best-fit judgement against the provider-level rating descriptors</td><td>Assessment by the pilot subject panels:<br>• Three-step method of assessment<br>• Best-fit judgement against the subject-level rating descriptors</td></tr>
<tr><td colspan="2">Outcomes</td><td>Provider-level rating and statement of findings</td><td>Subject-level rating and statement of findings</td></tr>
</table>

**Figure 2: Pilot process**




31. A summary of the changes from the first pilot that were tested in the second can be found in Table 2.

**Table 2: Summary of key refinements made for the second pilot**

| Element | Summary of change |
|---|---|
| **Subject classification system** | The Higher Education Statistics Agency (HESA) Common Aggregation Hierarchy at level 2 (CAH2) was used with some refinements. |
| **Framework** | Refinements to criteria:<br>• Separated the TEF criterion 'TQ1: Student engagement' into two distinct criteria<br>• Distinct criteria and rating descriptors at each level.<br>Refinements to the evidence:<br>• Expanded the range of contextual information<br>• Introduced new National Student Survey (NSS) metrics on learning resources and student voice (each half weighted)<br>• Tested a different combination of core metrics relating to student outcomes, drawn from Longitudinal Education Outcomes (LEO) and Destination of Leavers from Higher Education (DLHE) data<br>• Tested feasibility of new data on differential degree attainment and contextual data on grade inflation |

| Element | Summary of change |
|---|---|
| | • Tested more directive guidance about the content of submissions<br>• Tested an approach to ensuring that students have meaningful opportunities for involvement in submission<br>• Tested the inclusion of a provider summary statement to be used in subject-level assessment<br>• Revised examples of evidence that can be used in provider and subject-level submissions.<br>Refinements to the assessment and outcomes:<br>• Revised formula for the starting point of the initial hypothesis<br>• Revised panel structure and processes, to test scalability for full subject-level TEF<br>• Revisions to panel membership and roles, to ensure student membership on the pilot main panel and to deploy widening participation (WP) expertise across subject panels<br>• Tested how to make the outcomes more informative and useful for enhancement, including statements of findings at subject level. |
| **Model of assessment** | Comprehensive assessment of all subjects in a revised model that included:<br>• Provider-level assessment following a similar model to the current provider-level TEF<br>• Subject-level assessment with contextual information, metrics, submissions and ratings for each CAH2 subject<br>• Distinct criteria and rating descriptors at each level.<br>We also tested how the assessment at each level could interact. |
| **Limitation of data at subject level** | We explored ways of mitigating limitations in the data at subject level, including:<br>• The application of minimum assessment thresholds for cohort sizes and number of data sources<br>• Approaches to maximising the use of available data<br>• How submission can best address data limitations<br>• Presentation of subjects that are not able to receive ratings. |
| **Interdisciplinarity** | We explored how far additional measures taken better accommodated interdisciplinary provision in subject-level assessment:<br>• Students continued to be counted pro rata in the subject-level metrics against each subject that their course is mapped to for all interdisciplinary provision |

| Element | Summary of change |
|---|---|
| | • Better contextual data at subject level that included information about interdisciplinary provisions was used<br>• Specialist interdisciplinary panel member roles were tested. |

## Evaluation activities

32. Throughout the pilot the OfS sought and responded to live, informal feedback from participants in order to refine the model, test process options and troubleshoot issues. Formative feedback sessions were deliberately built in throughout: providers and student representatives attended a series of briefing events and workshops which incorporated feedback sessions. Similarly, all panel meetings (including training, calibration, and the assessment meetings themselves) had dedicated time set aside to gather feedback.

33. As with the first pilot, we conducted this pilot's activities with a greater risk appetite than we would in a real exercise and anticipated that challenges would arise. We developed and rolled out some novel processes in short time frames, often making adjustments in real time that relied on the good will and onward dissemination of information by multiple parties.

34. Alongside the panel meetings we established a programme of observations, where the main panel chair and deputies and senior OfS officers rotated attendance across all subject panels, with a view to supporting consistent practices across panels and to understand whether assessment issues or challenges arising were isolated or more systematic.

35. The OfS used three methods for formal evaluation of the design, process and implementation of the second pilot: provider surveys, panel feedback reports and OfS data analysis.

### Provider feedback

36. The OfS contracted IFF Research (IFF) to evaluate participating universities' and colleges' experience and perceptions of the pilot. IFF carried out their research in four strands: an online survey exploring the TEF processes; workshops with TEF contacts and student representatives; in-depth interviews with TEF contacts and academic leads; and an online survey focused on reactions to the outcomes and statements of findings. The research gathered responses from four stakeholder groups:

    a. The TEF main contact (the nominated OfS contact for an institution who oversaw the subject-level TEF submission process)

    b. The TEF main student representative (such as a students' union officer or sabbatical officer) nominated by an institution to be involved and contribute to the process, and to complete the student declaration

    c. Academic contributors, who led on the writing of subject submissions in their department

    d. Student contributors, who were engaged in the process and could contribute to the process at subject level.

## Panel feedback

37. The chairs and deputy chairs of each panel, as well as the student deputy chair and the WP and employment expert members of the main panel wrote reports on behalf of the panels. They reflect feedback gathered throughout the process and at a set of final meetings focused on evaluation. These reports are included as annexes to this report:[15]

    - Annex A: Main panel chair's report
    - Annex B: Student findings report
    - Annex C: Subject panel reports
    - Annex D: Widening participation expert report
    - Annex E: Employment expert report.

## OfS data analysis

38. The OfS undertook data analysis throughout the exercise and on the final ratings. This is presented in the 'Summary ratings and analysis' section of this report. Full details are presented as annexes:

    - Annex F: TEF subject-level pilot analysis
    - Annex G: Logistic regression models
    - Annex H: Type I and II errors in TEF.

[15] All annexes to this report are available at: www.officeforstudents.org.uk/publications/tef-findings-from-the-second-subject-level-pilot-2018-19/.

## Provider recruitment

39. In total 45 higher education providers took part in the second subject-level pilot. The full list of participants can be found in Table 4.

40. Providers were invited to submit expressions of interest for participation in the pilot in September 2018. This deadline was prior to the publication of the government response to its consultation and the pilot guidance, therefore providers were not aware of the details of the model to be piloted or how much resource would be required to participate when expressing interest. Following the release of the government response and pilot guidance, providers were better able to understand participation requirements and were asked to formally confirm their interest in participation. A small number of providers withdrew both at the confirmation stage and later in the process after the pilot started (these providers are not included in the 45 that completed the exercise). However, the sample of participating providers continued to reflect the diversity of providers in the higher education sector (see Table 5).

41. Purely for the purpose of this report we have categorised providers into three broad groups.[16] These are not formal categories used by the OfS for any other purpose:

    a. Further education colleges (FECs)

    b. Providers with more than 10 subjects assessed in the pilot (excluding FECs), broadly comprising large multi-faculty universities

    c. Providers with fewer than 10 subjects assessed in the pilot (excluding FECs), broadly comprising smaller and specialist providers, and those formerly referred to as 'alternative providers'.

42. We deliberately sought to recruit a number of single-subject providers to the pilot. Single-subject providers fall into two categories:

    a. 'True' single-subject providers, that only deliver provision in a single CAH2 subject

    b. Providers that are single-subject 'in effect', as they deliver provision across more than one CAH2 subject, but only have one subject with sufficient data to meet the minimum requirements to be assessed in the pilot.

    Two 'true' single-subject providers, and two 'in effect' single-subject providers were assessed in the pilot.

43. For 'true' single-subject providers, the data underlying the subject-level workbooks was identical to that underlying the provider-level workbooks. It was therefore decided not to duplicate assessment across both levels, and a single assessment was completed which determined a single rating for both the subject, and the provider as a whole. Providers involved in the pilot expressed a strong preference to be assessed by subject panels, so the

[16] It should be noted that IFF's research uses a different approach, which categorises providers into four broad groups which are closely aligned, but not identical, to the categories used in this report: Further Education Colleges (FECs), University, Specialist University, and alternative providers (APs). Although these categories formed the historic basis for the regulation of higher education providers, the OfS no longer distinguishes between provider types in this way. The alternative groupings used in this report are intended to more accurately group together providers that are likely to share common characteristics.

subject-level evidence base was used, and subject-level assessment processes were followed to determine outcomes for these providers.

44. For 'in effect' single-subject providers, provider-level data and subject-level data was not the same, as the provider-level metrics include aggregated data for the subjects which did not meet the threshold for assessment. In these cases separate assessments were completed at provider and subject level, with the expectation that different outcomes may be reached by each assessment.

**Table 4: Participating providers by type**

| Further education colleges | Providers with 10 or more subjects assessed in the pilot | Providers with fewer than 10 subjects assessed in the pilot |
|---|---|---|
| Bath College | Aston University | Academy of Contemporary Music Limited |
| Bishop Burton College | Birmingham City University | BIMM Limited |
| Burton and South Derbyshire College | Cardiff Metropolitan University | Bishop Grosseteste University |
| Chichester College Group | Goldsmiths' College | Falmouth University |
| City of Sunderland College | King's College London | Kaplan Open Learning (Essex) Limited |
| Gateshead College | Liverpool John Moores University | Rose Bruford College of Theatre and Performance |
| Hugh Baird College | Solent University | Pearson College Limited |
| New College Durham | Swansea University | Nelson College London Limited |
| The Oldham College | Teesside University | The University of Buckingham |
| South Thames Colleges Group | University of Central Lancashire | The University of Law Limited |
| The Trafford College Group | University College London | |
| Warwickshire College | The University of East Anglia | |
| | University of Lincoln | |
| | The University of Leeds | |
| | The University of Nottingham | |
| | The University of Reading | |
| | University of Suffolk | |
| | University of Wales Trinity Saint David | |
| | The University of West London | |
| | The University of Westminster | |
| | University of Wolverhampton | |
| | University of Worcester | |
| | University of York | |

**Table 5: Participation in TEF subject-level pilot 2018-19 by provider type compared to proportion of UK higher education sector**

| | Sector | | Pilot | | |
|---|---|---|---|---|---|
| | % providers | % students | Number of providers | % providers | % students |
| **FEC** | 48 | 5 | 12 | 27 | 2 |
| **Not FEC, 10 or more subjects assessed** | 22 | 87 | 23 | 51 | 92 |
| **Not FEC, less than 10 subjects assessed** | 30 | 8 | 10 | 22 | 6 |

**Table 6: Participation in TEF subject-level pilot 2018-19 by provider size compared to proportion of UK higher education sector[17]**

| | Sector | | Pilot | | |
|---|---|---|---|---|---|
| Size of provider | % providers | % students | Number of providers | % providers | % students |
| **Fewer than 500 students** | 51 | 3 | 8 | 33 | 1 |
| **500 – 4,999 students** | 27 | 10 | 15 | 49 | 7 |
| **5,000 or more students** | 22 | 87 | 22 | 18 | 92 |

**Table 7: Participation in TEF subject-level pilot 2018-19 by provider tariff compared to proportion of UK higher education sector**

| | Sector | | Pilot | | |
|---|---|---|---|---|---|
| Tariff | % providers | % students | Number of providers | % providers | % students |
| **High** | 6 | 22 | 5 | 11 | 29 |
| **Medium** | 14 | 27 | 11 | 24 | 33 |
| **Low** | 70 | 37 | 26 | 58 | 27 |
| **Mixed** | 6 | 13 | 3 | 7 | 11 |
| **N/A** | 3 | 0 | 0 | 0 | 0 |

[17] Calculated using full-person equivalent, which is a headcount measure. The concept of full-person equivalent student numbers is defined in full at: https://www.hesa.ac.uk/support/definitions/students.

**Table 8: Distribution of TEF ratings for the sector versus the pilot sample[18]**

| | Sector | | Pilot | |
|---|---|---|---|---|
| **Final rating** | **Number of providers** | **% providers** | **Number of providers** | **% providers** |
| **Bronze** | 61 | 22 | 17 | 38 |
| **Silver** | 135 | 50 | 16 | 36 |
| **Gold** | 76 | 28 | 11 | 24 |
| **No rating** | - | - | 1 | 2 |

[18] Calculated from TEF provider ratings on 27 August 2019 (See: www.officeforstudents.org.uk/advice-and-guidance/teaching/tef-outcomes/#/tefoutcomes/). Analysis does not include Provisional ratings.

## Submissions

**There was a clear expectation by staff and students that subject-level TEF would help drive improvements in higher education.** Staff and students in participating providers reported that they were already seeing positive impacts, and they expected these benefits to increase over time. They reported senior buy-in and saw opportunities for subject-level TEF to bolster internal enhancement processes.

**Participation in the pilot was resource intensive for providers but there is scope to integrate TEF with internal processes.** It took considerable time, effort, co-ordination and mobilisation of resources for providers to understand and interpret their subject data and to pull together the range of submissions required. However, there was a strong sense that subject-level TEF could become more aligned with internal processes, and recognition that the process would become more efficient once it is familiar and embedded. This would be aided by the OfS ensuring that a future iteration of TEF is well aligned with regulatory mechanisms, in particular access and participation plans and ongoing monitoring against the conditions of registration.

**Students at most providers had opportunities to engage meaningfully in submissions, and there is scope to extend their contribution further.** Despite limited time and other constraints, the lead student representatives at most providers were positive about students' opportunities to engage in TEF pilot submissions. They suggested that the TEF in future could go further in including evidence more directly from students. Many providers proactively encouraged and facilitated student engagement in the TEF, but they reported that engaging students in TEF is harder at subject level than at provider level.

**Time pressures adversely impacted on metrics production and submissions.** Due to timing pressures and newness of some aspects of the data, we encountered errors in the presentation of the data we released to providers taking part in the pilot and had to re-release it. This hampered their ability to prepare their submissions. In a full-scale exercise, the OfS would also need more time to process and quality assure data prior to release, and providers would need more time to analyse their data and prepare their submissions.

**Different types of providers reported differences in their experiences**. Larger and more established providers tended to be better equipped to analyse complex data, draw on additional internal data, engage with student representatives, and see opportunities to align the TEF in future with internal processes. FECs and alternative providers tended to find these processes more challenging, but some also saw subject-level TEF as an opportunity to compete with 'top ranking' providers, by demonstrating excellence in particular subjects.

45. In October 2018 we confirmed which providers would participate and published the pilot guidance and technical guidance. In a change to the previous TEF exercises, the OfS guidance incorporated both the specification (detailing the framework and evidence to be used, which was previously published by the DfE) and procedural information for participants.

46. Subject-level metrics were initially released to participating providers at the beginning of November 2018 along with online training resources, though, as described below, production errors meant that metrics were subsequently reissued.

47. Briefing events for providers and lead student representatives were held in mid-November 2018. This was followed by the release of additional training resources in December 2018 and further briefing and feedback events in January and March 2019 for TEF contacts and lead student representatives.

48. At the briefing and feedback events we focused on clarifying procedures and how to interpret metrics, write submissions, deal with missing data, and engage students in the process. These sessions were collaborative in spirit and were supported by pilot participants who gave presentations on the above topics. Whilst we gathered some feedback on the framework itself and early views emerged, much of the discussion was process-focused.

49. Outside of the events, the OfS provided a general helpdesk and a metrics helpdesk. Queries peaked once providers could access their metrics, typically focusing on metrics definitions, scope and coverage. The majority of queries related to subject mapping and clarifying or amending the subjects that were in and out of scope for assessment.

50. A subset of pilot participants who had participated in the first pilot were asked to submit some of their submissions three weeks early to provide calibration materials to support panel member training. The remainder of the submissions were received by the deadline at the end of February, though there were several providers who required short extensions.

51. Feedback was not given to providers who submitted early for the calibration exercise, but, recognising time limitations, they were able to resubmit revised submissions at the normal deadline if they wished to.

### Identifying which subjects would be assessed

52. The OfS initially identified 744 subjects that were potentially in scope across all pilot providers. Of these, 645 subjects were pre-determined by the OfS to be both in scope and also to meet the definition of sufficient data for assessment (two of the three 'metric types' and more than 20 students in the contextual data population for the majority mode of study). Providers had the opportunity to review which of their subjects were in scope for the pilot prior to writing submissions. Through this process, a further 15 subjects were agreed to be out of scope for assessment. Where a subject was agreed as out of scope, no submission was required and no subject rating generated, but data associated with the subject continued to contribute to the overall provider-level data.

53. There were a range of reasons why providers made a strong case for taking a subject out-of-scope, even though it met the definition of sufficient data for assessment. These included where a provider was no longer recruiting to any courses in that subject, and a number of examples where a course was mistakenly coded in one year to a subject that the provider did not offer (therefore creating sufficient metrics for that subject) but in previous and/or subsequent years was coded to its correct subject. Some requests were also rejected by the OfS. If scaling up the exercise, we would need more detailed and explicit criteria, and a more formal process, for deciding which subjects are in- and out-of-scope for subject-level assessment.

54. Whilst providers were able to report other concerns about data errors for future correction, the pilot timetable did not allow for a specific data amendments process. Some data amendments were incorporated into metrics released to the wider sector, following

completion of the separate Year Four TEF process for correcting errors in providers' data returns.

## Metrics releases

55. Generating the full set of metrics was complex and time-pressured. Production and release took longer than planned on account of complexities encountered in the development, testing and implementation of adequate data suppression strategies to meet data protection requirements. Following release, we also identified errors in the calculation and highlighting of absolute value markers in the workbooks which required correcting.

56. Many of the corrections only resulted in minimal changes, and did not result in changes to flags or initial hypotheses. Nonetheless, it was not until mid-December 2018 that all workbooks were correctly released. To ensure that providers were not disadvantaged by the delays, extensions were offered to a proportion of subjects due to the initial delay. We also used the fact that panels were due to assess material in two batches to accommodate some further extensions following the re-releases.

57. Whilst we were able to mitigate disruption to the overall assessment process, there was an impact on providers. Unsurprisingly, when surveyed by IFF the majority of TEF main contacts felt that data was not provided to them in a timely manner (54 per cent) and negative feedback reflected the difficulties in releasing the metrics.

> "We've got a team of people doing a lot of work on these things. We were encouraged to, as we tried to in the narrative, shape the narrative around to explain, to mitigate, where relevant, the metrics, and then to get them re-released is really unhelpful."
>
> TEF main contact (University)
> TEF subject-level pilot evaluation – Provider perspectives

58. In early February 2019 the OfS released pilot metrics to all providers in the UK. These resources were made available to help providers develop their approach to TEF at subject level and respond to the independent review's call for views.

## Views from providers and student representatives

59. As noted above, briefing events offered an opportunity for the OfS to gather informal feedback ahead of the formal IFF surveys. A number of early themes emerged:

    a. Providers suggested that the metrics could be simplified to make the submission process less burdensome. An example of complexity was the difference between, and interaction of, the double and single flags and absolute value markers. Other concerns included the weighting of metrics and various issues relating to LEO.

    b. Providers were concerned about tight timescales. In particular, they would have liked more time to consider the metrics and produce submissions. The timetable could have been more sensitive to how submission deadlines interacted with academic timings such as exam periods and the period during which the National Student Survey (NSS) is open for responses.

c. Providers wanted a greater level of student involvement and for the student voice to be meaningful in the subject-level process. But many found it challenging - both TEF contacts and student representatives would have welcomed more time and guidance on good practice.

d. Providers welcomed the improvements made to the subject categories, but reported that challenges remained when their internal structures did not map well to the categories. A number of providers suggested there should be flexibility for each provider to tailor the subject categories to achieve a better fit with their provision.

e. Some providers felt that a future iteration of TEF should be more closely linked to other regulatory and enhancement mechanisms, in particular access and participation plans and ongoing monitoring against the conditions of registration.

60. These early themes carried through to the independent surveys carried out on behalf of the OfS by IFF Research (see Box 1 below).

### Box 1: Provider perspectives

**This box reproduces the key findings from IFF's executive summary. The full report can be found on the OfS website**.[19]

The [IFF] study aimed to capture evidence about the experience of higher education providers taking part in the second year TEF subject-level pilot. The research was designed to address a number of questions, which can be broadly categorised into two themes of 'process' and 'outcomes'.

Under process, the research explored how institutions incorporated student voice into their TEF submission, the extent to which they found the evidence and assessment procedures to be robust and generally what burden the submission process placed on the institution. In terms of outcomes, key questions were asked about the current and potential future impact of subject-level TEF should it be rolled out across the higher education sector, including the extent to which it will drive enhancements in teaching and learning, support diversity of provision and widening participation and social mobility.

#### Process

##### Meaningful input from students

Student voice was represented through two key roles: student representatives and student contributors. **Just over half of student representatives (56 per cent) said the role had provided a meaningful opportunity for students to engage with the process.** Qualitative feedback found a lack of clarity around the declaration form and its objectives, with several students commenting that it needs to be less of a tick-box exercise.

Engaging students more widely was challenging. **Around one in three (36 per cent) of TEF main contacts said that it had been difficult to engage student representatives, and nearly double this proportion (65 per cent) cited difficulty engaging student contributors.**

[19] The report 'TEF subject-level pilot evaluation - Provider perspectives' is available at: www.officeforstudents.org.uk/publications/tef-findings-from-the-second-subject-level-pilot-2018-19/.

Undoubtedly the timing of the process did hamper efforts to engage students. **Where providers did have more success engaging students, this appeared to stem from proactive attempts made by academic staff to communicate with and encourage their students, as opposed to approaches from other departments or students.**

### Robust evidence and assessment processes

**The subject-level TEF metrics were, on the whole, found to be complex**. Larger, more established, institutions were better prepared and able to handle the complexity of the exercise, often identifying a specific individual who was tasked with focusing on the data analysis before sharing that insight with other contributors. **A minority of respondents spoke positively about the data and the insights they gleaned from it.**

**Measures from NSS, including teaching on my course, assessment and feedback, and academic support, were seen as most relevant**. In contrast, LEO data on sustained employment/ further study and median earnings/ higher study were seen as least relevant.

**There was broad support across all the assessment criteria.** Feedback, resources and rigour/ stretch were seen as most relevant.

**Providers had mixed views on the statement of findings and their usefulness**. Some said the narrative contribution helped to provide context around the metrics, whereas others felt that the narrative explanation they offered was overshadowed by the data and not taken into account by the panel. Some commented on perceived inconsistencies across subject areas in the decision-making; others said it was difficult to disaggregate learnings by subject area given the groupings and others struggled to marry inconsistencies between provider-level and subject-level ratings.

### Institutional burden

A number of providers remarked on the length of time it took to prepare their subject-level TEF submission. **Over half (56 per cent) of academic contributors spent at least a week contributing to their institution's subject-level TEF process; 25 per cent spent at least a fortnight.**

## Outcomes

### Driving enhancement

In terms of current impact, **46 per cent of student contributors said that subject-level TEF has already had a positive effect on the learning environment, with 42 per cent saying there has been a positive effect on teaching quality and widening participation.**

31 per cent of academic contributors reported a positive impact of subject-level TEF on teaching quality and activities, while 25 per cent reported a positive impact on the learning environment. **Over half (58 per cent) also stated more broadly that subject-level TEF would act as a tool for internal enhancement leading to continuous improvement**.

### Influencing prospective student choice

**Student contributors demonstrated a preference for subject-level TEF over provider-level TEF, with 89 per cent saying that it is more useful than provider-level TEF. More than half of student representatives (56 per cent) agreed.** Asked to explain their reasoning, students often focused on subject as being the primary concern for prospective students ahead of where to study. Staff were more sceptical about the potential for subject-level TEF to influence the choice of prospective students.

### Supporting diversity of provision

There was some scepticism as to whether TEF recognises diversity and innovative forms of excellence across a diverse sector. **Only 23 per cent of academic contributors and 14 per cent of TEF main contacts agreed that subject-level TEF would – in its current form – support diversity of provision.**

### Supporting widening participation and social mobility

Staff and students were broadly split as to whether subject-level TEF will support widening participation and social mobility. **Around a third of staff (36 per cent TEF main contacts, 31 per cent academic contributors) agreed that subject-level TEF would have a positive impact on supporting widening participation and social mobility**. Meanwhile, around two-fifths of student representatives and contributors felt there had **already** been a positive effect as a result of subject-level TEF on widening participation.

### Effects on provider behaviour

There were a number of points where the research focused on learnings from the pilot and how higher education providers might do things differently in the future or what they would do to prepare for subject-level TEF should it be rolled out. **The responses – particularly among more established higher education providers – typically focused around efforts to better align the subject-level TEF process with provider's own existing internal quality assurance and enhancement processes, thus reducing the aforementioned level of burden and complexity.**

## Student engagement in submissions

61. Student engagement in the TEF submission process has been promoted in previous TEF exercises, recognising the additional insight students can provide. In previous exercises providers have been encouraged to share their TEF metrics with student representatives and engage them in writing submissions.

62. Through the pilots, we explored how to strengthen student engagement overall, and tested how it could work at subject level. Whilst we encouraged providers in the first pilot to involve students, feedback from students indicated they would like a more formal mechanism to ensure that students have good opportunities to be involved in, and contribute to, the submissions should they wish to.

63. The guidance for this pilot set out clear expectations for providers to seek to engage students in the process at both provider and subject levels. This could be through, for

example: the use of surveys, representative structures, focus groups, student representation on relevant committees, consultation events, facilitating the students' union to draft a section of the provider submission or include a supporting statement of endorsement.

64. To test a way of ensuring that providers would seek to involve students, and to support students in doing so, we introduced:

    a. **TEF lead student representatives**. Providers were asked to nominate a student representative to:

        - be involved in the provider submission should they wish to be
        - access the pilot TEF metrics to support their involvement in the submissions
        - help co-ordinate student involvement in subject-level submissions
        - attend OfS pilot briefing events
        - provide evaluation and feedback to the OfS.

    b. **The student declaration**. The TEF lead student representative was also asked to complete a declaration to indicate whether and how students were involved in submissions. These declarations were submitted directly to the OfS and were considered by the pilot main panel as additional contextual information.

65. Due to the varied nature of student representation across the sector, the OfS was not prescriptive about who a provider should nominate as their lead student representative, although we expected them to use existing structures of student representation.

66. The declarations were completed by the TEF lead student representatives, who were responsible for coordinating feedback from students at their provider. They were asked to fill out a pro-forma with a factual summary of how students were involved in the pilot submissions and whether students had meaningful opportunities for involvement. In answering these questions, lead student representatives were instructed to try and represent the views of the wider student body.

67. Guidance to lead student representatives acknowledged that there would be different approaches to good student engagement but was clear that the aim was for students and their university or college to have entered into an authentic partnership to jointly engage in the TEF pilot. It was also recognised that this partnership might not have been possible in practice. There was no obligation for the TEF lead student representative to share this declaration with their university or college.

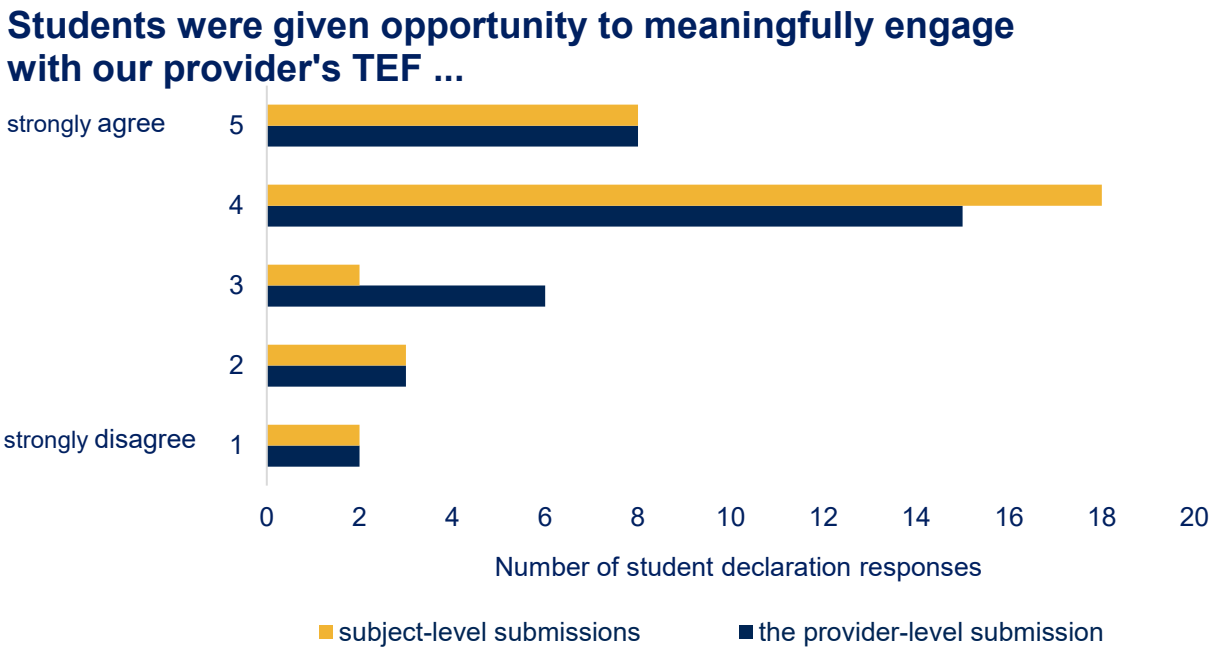
68. Student representatives also had the opportunity in the declaration to state that they did not want to engage in the process, and provide supporting comments about this if they wished to.

69. A total of 37 student declarations were submitted by TEF lead student representatives in subject pilot providers and it is notable that FECs represented five out of the eight non-

submissions, reflecting the differing levels of student representation structures across the sector. The student body at one provider opted not to engage in the process.

70. Whilst the majority of declarations were positive, a small minority of respondents indicated that they had not had meaningful opportunities to engage with submissions, as Figure 3 illustrates.

**Figure 3: Student declarations about their engagement in provider-level and subject-level submissions**




71. The TEF subject-level pilot evaluation – Provider perspectives[20] report found that:

- Just over half (56 per cent) of student representatives felt that the student declaration provided a meaningful opportunity to engage with subject-level TEF.
- The majority of student representatives (72 per cent) found it easy to complete; in contrast one in four (24 per cent) said they found it difficult.
- In terms of improving the declaration, the most common theme mentioned by half of student representatives who responded, was to clarify the objectives of the student declaration and make sure that that questions are not repetitive.
- Student representatives were interested in alternatives to the student declaration. The option of a student-authored section in the provider-level TEF submission was most popular, gaining the support of just over half of student representatives (52 per cent).

[20] TEF subject-level pilot evaluation report - Provider perspectives. Available at: www.officeforstudents.org.uk/publications/tef-findings-from-the-second-subject-level-pilot-2018-19/.

"The student declaration needs to be clearer in what it is asking… the questions were very vague and this didn't help when filling the declaration out".

Student representative (Further Education College)
TEF subject-level pilot evaluation – Provider perspectives

"Less vague questions, actually change it to an interview."

Student representative (University)
TEF subject-level pilot evaluation – Provider perspectives

72. Whilst some very positive student engagement was observed in the pilot, the non-submission of some student declarations and low scores in others raises concerns about how scalable these processes would be for providers. A clearer purpose for the student declaration and enhanced engagement, support and guidance for student representatives and contributors would help, but further work, building on the OfS's wider work to develop student engagement[21], would need to be done to ensure it is meaningful and adds value.

[21] See: www.officeforstudents.org.uk/advice-and-guidance/promoting-equal-opportunities/effective-practice/student-engagement-and-consultation/.

# Panel structure, recruitment and training

**The panel members we recruited were of a very high calibre and we achieved the range of expertise needed for the pilot, but expanding the panels for a full subject-level exercise would present challenges.** Three to four times as many panel members would be required for a full exercise and it could be difficult to maintain the level of expertise of pilot panels. Certain types of providers were underrepresented on the pilot panels and it would be challenging to overcome this with larger panels. It would also be important to improve the diversity of panel members in terms of their personal characteristics.

**Further improvements to, and more time for, panel training and panel calibration activities would be required in a scaled-up exercise.** Effective training and panel calibration have been crucial to enable robust and consistent assessments. In the pilot we trialled a mix of online and face-to-face training. This approach was welcomed but further improvements, tailoring of training and more time would be needed to prepare effectively a much larger and more diverse range of panel members for a full subject-level exercise.

## The pilot panels

73. 147 panel members carried out the pilot assessments. They were selected for their standing in the higher education sector, breadth of expertise, and commitment to excellence in teaching.

74. The pilot main panel was responsible for deciding the provider-level ratings, and for checking for consistency across the subject panels. The main panel comprised:

    a. the main panel chair

    b. the deputy academic chair

    c. the deputy student chair (the second pilot looked to enhance student engagement throughout, and as part of this a new deputy chair role on the main panel was created specifically to be held by a student)

    d. the co-chairs and student deputies from the subject panels

    e. additional academic members and additional student members

    f. widening participation (WP) and employment experts.

75. As full subject-level TEF would require subject panels to make thousands of subject-level assessments, we configured pilot subject panels to test approaches that could be scaled up to a full subject-level exercise. To do this, we had five ‘paired’ subject panels in the pilot. Each of the five ‘paired’ pilot subject panels comprised two main clusters of subjects (as shown in Table 9) but with the view that they could operate as two separate subject panels in future. Each subject panel included:

a. An academic chair and a student deputy chair who were also members of the main panel.

b. Student representatives and academics (with one student for every two academics), and additional representatives from employers, and professional, statutory and regulatory bodies (PSRBs).

c. A student or academic member of each panel who was identified to act as the interdisciplinary liaison. This member had additional responsibilities to advise the subject panel on matters relating to interdisciplinary subjects.

d. A student or academic member of each panel who was identified to act as a widening participation liaison. This member had additional responsibilities to support the panel by highlighting widening participation issues and working alongside experts on the main panel.

**Table 9: Combined subject panels**

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

## Panel recruitment

76. Panel recruitment consisted of:

a. matching and re-appointment of chairs and deputies from the first pilot

b. matching and re-appointment of panel members from the first pilot

c. open recruitment of vacant roles.

77. 147 panel members carried out the pilot assessment process. Most panel members were appointed in the first year of the pilot (2017-18), following an open recruitment process. Panel members were selected against published criteria, such as their commitment to excellence in teaching, standing in the higher education sector, professional expertise (academic panellists) and experience of student representation (student panellists). The process aimed to represent the diversity of the UK higher education sector across all panels, for example ensuring a mix of panel members from different types of providers, geographic locations and subject specialisms.

78. Recruitment was phased to first appoint chairs and deputies who could then be involved in finalising the make-up of their subject panels. All 14 subject panel chairs and deputy chairs from the first pilot were re-appointed and matched to the panel best matching their subject

expertise. Additionally, three academic co-chairs, and three student deputy chairs were recruited to fill vacancies which arose as a result of the panel restructuring. These positions were as follows:

a. Medical Sciences panel – academic co-chair and student deputy

b. Social Sciences panel – academic co-chair

c. Education and Social Care panel – academic co-chair

d. Education and Social Care, and Natural and Built Environment panels – student deputy chairs.

79. The distribution of subject expertise required for the second pilot meant that the vast majority of panel members were able to be retained from the first pilot. In the second pilot, recruitment was limited to filling specific gaps in subject expertise, provider type representation or specialist roles such as employer representatives. These vacancies were advertised for open recruitment. Targeted communications were aimed at specific subject groups, FECs and student groups. In particular, we undertook work to engage with: FECs (via the Association of Colleges (AoC) and The Mixed Economy Group (MEG) of colleges); employers (Confederation of British Industry (CBI), Federation of Small Businesses (FSB), and Institute of Student Employers (ISE)); and professional, statutory and regulatory bodies (PSRBs). Particular organisations were targeted where there was a specific need on the panels, and networks were used wherever possible to promote the vacancies.

80. In some subject areas, recruitment proved particularly challenging. For example, we needed to extend the recruitment window to recruit academic panel members from further education colleges and subject experts in Nursing and Allied Health and Health and Social Care. Some PSRB and employer positions were also difficult to fill, for example where the employment route for a discipline was not clear there was no obvious representative body. For the pilot, we asked existing panel members to suggest individuals and organisations that might be suitable; however, this approach might not be scalable. Significant targeted recruitment activity would be required to ensure full representation of expertise across all panels in a full-scale exercise.

81. We requested equality protected characteristics monitoring information from applicants during the application process in 2017-18 (TEF Year Three and the first subject pilot) and sought applicant permission to use their diversity data in any tie-break situations (a legally permitted form of positive action aimed at increasing panel diversity).[22] We compared the diversity profile of applicants and appointed panel members against senior staff sector averages, to check for bias in our selection process.[23] As the majority of panel members in

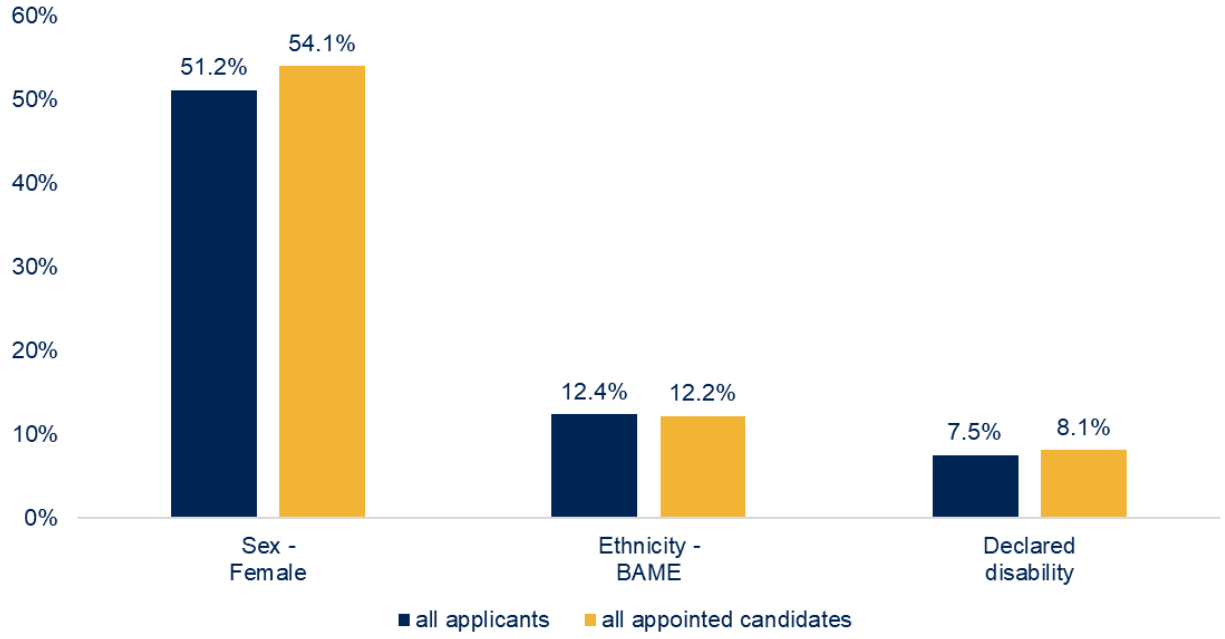
[22] Monitoring by self-declaration using census monitoring categories. BAME included all black, Asian and mixed ethnicity backgrounds.

[23] We compared academic applications and appointments data against sector data for senior staff. Source: HEFCE analysis of HESA 2015/16 data for 'Academic leadership' and 'Professors' (excluding atypical staff) at HEFCE-funded higher education institutions.
See: https://webarchive.nationalarchives.gov.uk/20180405122231/http://www.hefce.ac.uk/analysis/eddata/.
We compared student applicant and appointee data against Office for Students' analysis of HESA data. Source: Student population - All students (all domiciles), Level of study - All levels of higher education, Student characteristic - Entire sector. See: www.officeforstudents.org.uk/data-and-analysis/equality-and-diversity/equality-and-diversity-data/.

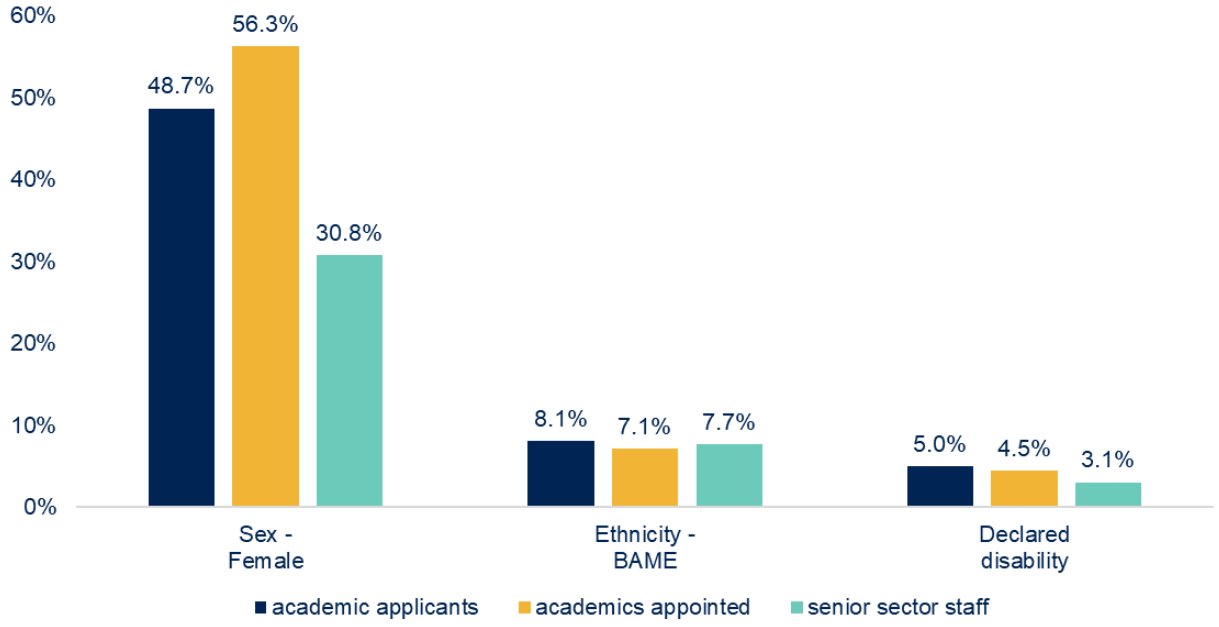
the second pilot continued in post from the first pilot, the diversity of the panels, when looking across all panels, remained broadly comparable across both years though naturally representation varied between panels.

82. Panellist monitoring data indicated that TEF panels were mainly in line with sector averages (see Figure 4), or more diverse than the sector particularly in terms of female academics (see Figure 5) and disabled students (see Figure 6).

**Figure 4: All panel applications and appointments by personal characteristics**

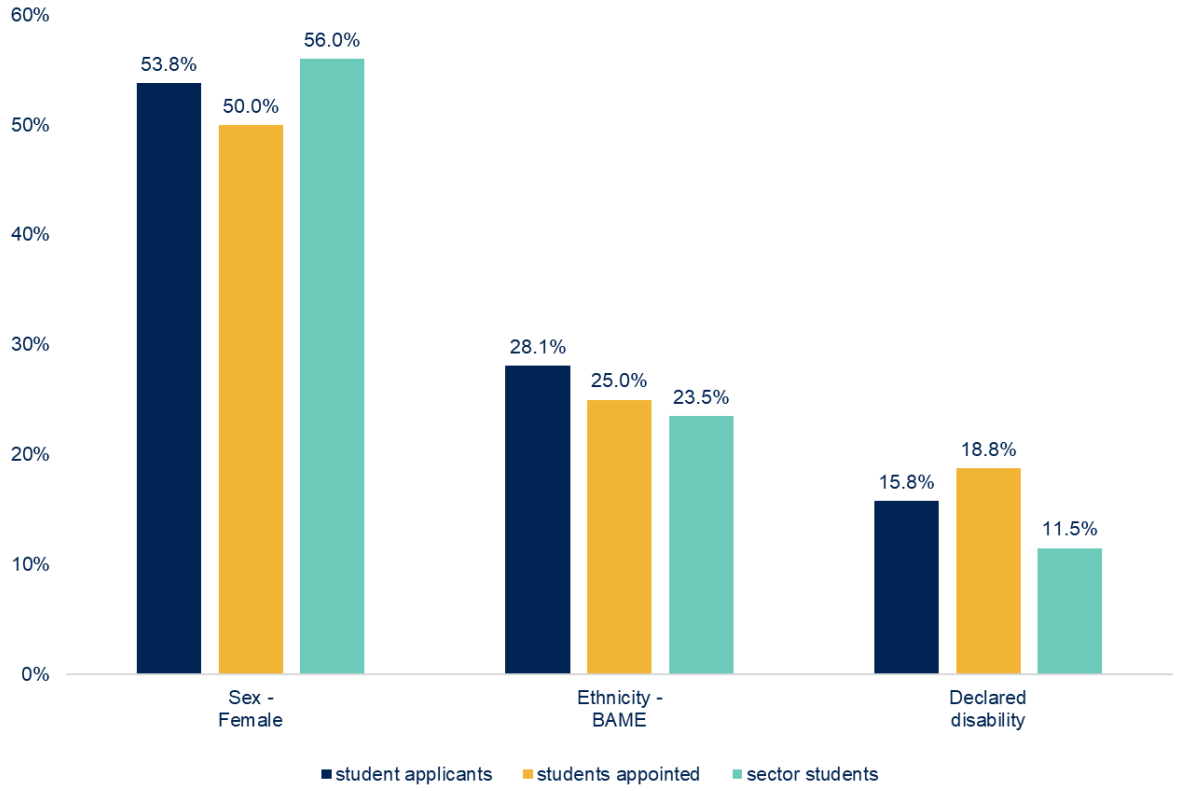


## Figure 5: Academic panel applications and appointments by personal characteristics[24]




Source: HEFCE analysis of HESA data[25]

## Figure 6: Student panel applications and appointments by personal characteristics




Source: Office for Students analysis of HESA data[26]

[24] HEFCE analysis of HESA 2015/16 data for 'Academic leadership' and 'Professors' (excluding atypical staff) at HEFCE-funded higher education institutions. See: https://webarchive.nationalarchives.gov.uk/20180405122231/http://www.hefce.ac.uk/analysis/eddata/.

[25] See note 24.

83. We recognise, however, that inequalities remain at a sector level and will continue to proactively promote equality, diversity and inclusion throughout the recruitment of panel members to increase panel diversity. We also acknowledge differences in the make-up of individual subject panels, and concerns raised by student panel members in particular around equality, diversity and inclusion issues.

## Training

84. All panel members were trained through a mixture of online videos, e-learning modules and face-to-face sessions. All panel members were enrolled in two online courses covering the technical aspects of the metrics and the steps of assessment. A main panel training event in January 2019 was attended by 36 panellists, followed by a subject panel training event attended by 124 panellists.

85. These mixed training methods represented a new approach taken in this pilot, which aimed to explore more scalable modes of delivery than relying solely on large in-person events. While training days had been sufficient in provider-level exercises, they had proved extremely challenging to deliver at the scale required in the first pilot, let alone the expected scale that would be required for a full subject-level exercise. The online training methods tested in this pilot were found to have a number of benefits in terms of cost-effectiveness, greater flexibility for panel members to complete training in their own time, and new opportunities to check understanding and engagement through online testing and learner analytics. However, greater time and resource would need to be devoted to development and testing of good quality online resources if this were to be more heavily relied upon in a real exercise. Online materials used in the pilot were limited to the generic assessment process and were not tailored for different panel roles, which was not explained until later in the process. Training software was selected based on its administrative simplicity and usability, which was appropriate for proof of concept in the pilot but did not meet the needs of pilot panel members.

## Calibration

86. Training was followed by main panel and subject panel calibration exercises. Each panel worked through the assessment of a small sample of five to six submissions, from providers that had been asked to make some of their submissions early for this purpose. Panel members reviewed the calibration cases and attended a one-day meeting to discuss their assessments and calibrate their judgements. Calibration offered panel members the first opportunity to put their training into use with real materials, and was intended to establish a collective understanding of the assessment process and thresholds between ratings, before panel members started on their individual allocation of cases. We also used the time to test some panel processes, including trialling ways to operationalise the statement of findings (SOF) writing processes. We made a number of changes and improvements to the final SOF writing guidance based on panel feedback.

87. While many of the pilot panel members brought experience from previous TEF exercises, newly recruited panel members participating for the first time found the learning curve

[26] OfS analysis of HESA data. Student population - All students (all domiciles), Level of study - All levels of higher education, Student characteristic - Entire sector. See: www.officeforstudents.org.uk/data-and-analysis/equality-and-diversity/equality-and-diversity-data/.

particularly steep. In addition, a number of panels reported that extra and more tailored support would have been beneficial. In a full-scale subject exercise, panels would need to be expanded to three or four times the number of members and many members would be new to the role. More extended training and extended calibration would be needed.

88. The main panel and student reports suggest that it would be helpful to cover, for example:

    a. Training for chairs

    b. Self-confidence, resilience and negotiating workshops

    c. Equality and diversity training

    d. Unconscious bias training

    e. Data literacy training.

And additionally, the OfS has identified there would be a need for:

    f. General orientation

    g. Support for students in contributing effectively to assessment meetings

    h. Earlier and more comprehensive training for specialist roles

    i. Moderation activities to be included in calibration

    j. Interpretation of qualitative data.

## Assessment process

**Fully assessing all subjects was a more robust model for producing subject-level ratings.** All subject-level ratings were awarded on the basis of a full assessment of metrics and a submission specific to that subject. This largely addressed the key concerns with the previous year's pilot models – awarding subject-level ratings with limited evidence or no direct assessment.

**It was challenging to carry out assessments consistently across the different subject panels; this challenge would be exacerbated in a full-scale subject TEF.** We tested approaches to moderating across the five subject panels with the aim of achieving consistent standards of judgement. While this achieved a broad level of consistency, we found some differences in approach that were not able to be resolved in the time available. It would be more challenging and time consuming to try to ensure consistency across a greater number of panels and panel members in a full subject-level TEF.

**We tested decision-making processes that were designed to accommodate the greater volumes of decisions that would be required at full scale. These processes would need further development.** The decision-making process used in provider-level TEF progresses through individual reviews, initial small group (trio) discussions, group recommendations to refine the rationale, and finally full panel decisions. This process has

worked well in other TEF exercises but is time-consuming. We therefore piloted a compressed version without trio discussions, but panels found it hampered the quality and inclusivity of assessment discussions. An alternative approach would need to be developed to allow trios to confer prior to the group recommendation stage. This might require a mechanism to enable several thousand separate trio discussions to take place, while avoiding the logistical challenges of doing so within meetings.

**Data on differential attainment is important and could be developed further into a metric.** The pilot tested how experimental 'supplementary' data, indicating gaps in degree attainment between groups of students, could be used in provider-level assessment. The main panel agreed that differential attainment could be considered within the TEF, as a key issue to be addressed in the sector. However, the panel found that to make use of the data in assessment, it would need to be developed further into a metric. Also, further clarity would be needed on how it relates to our regulation of access and participation.

**Panels found student input to submissions was valuable where present and could be extended further.** The introduction of the student declaration, explaining how students were involved in the submission process, was a positive step. There was enthusiasm for the process to be extended further in future, with students more directly providing evidence to the panels. In developing this process, some tensions would need to be addressed, for example if student-submitted information is absent or contradicts the provider submission.

## Allocation and stage 1 processes

89. The process of assessment involved:

- Training and calibration exercises for all panel members to prepare them for assessment
- Each panel was divided into two groups of nine members; the submissions within each panel's remit were assigned to a group of nine and then allocated to three members within that group for detailed review
- The assessment proceeded as follows:

  **Stage 1**: Detailed individual assessment of each case by the three members.

  **Stage 2**: Discussion of each case by the group of nine members, to form a recommendation to the whole panel. Where appropriate, a fourth reviewer was identified.

  **Stage 3**: Consideration of all recommendations, and final decisions by the panel, with particular attention given to borderline or more challenging cases.

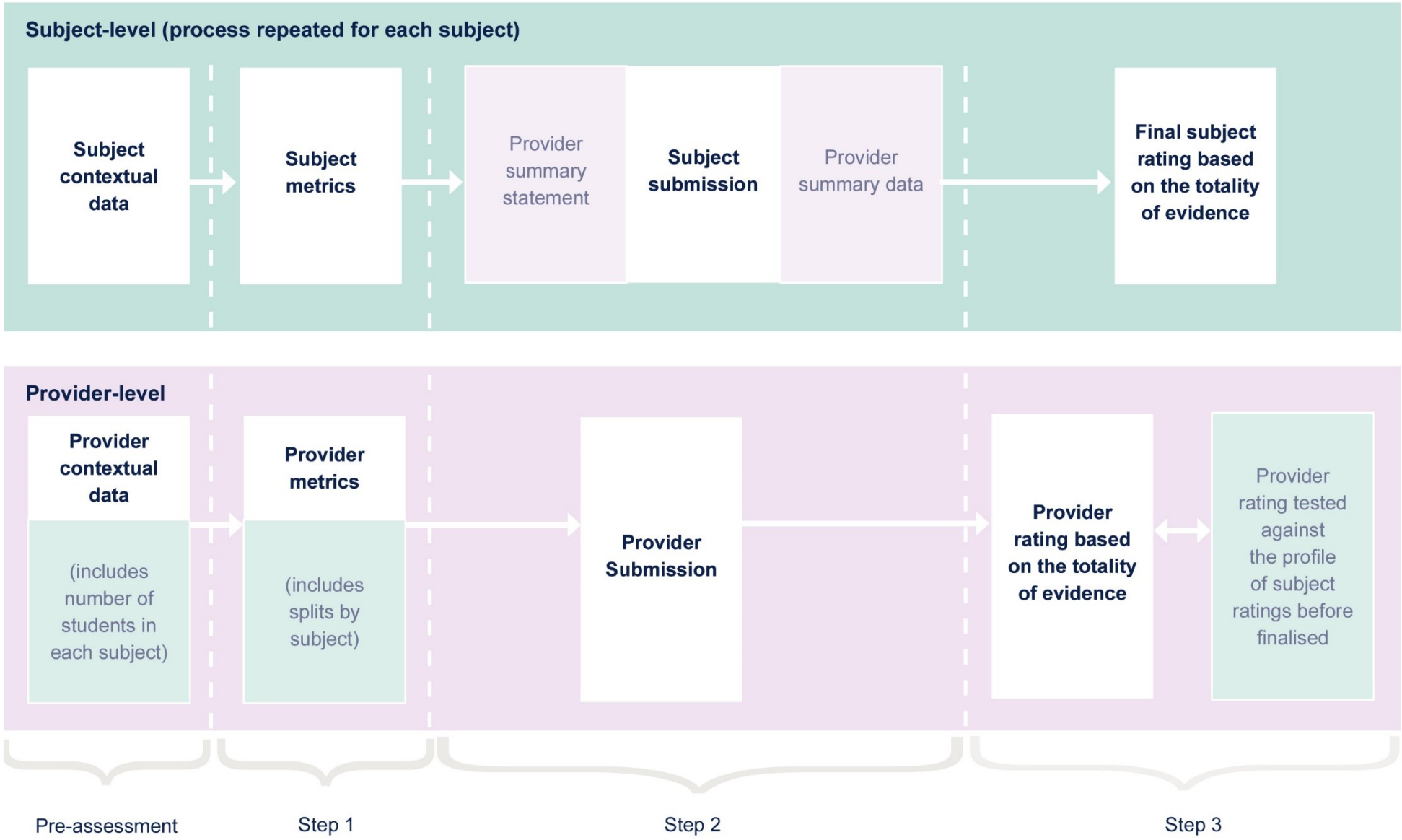
90. To manage the volume of subject assessments in the pilot, the submissions were split into two batches and the process outlined above was repeated for each batch, taking place sequentially across two sets of panel meetings.

91. Throughout the assessment, panel members followed a three-step method of reviewing the available evidence and arriving at a rating. This was adapted from the existing provider-level

TEF method in which an initial hypothesis is based on a review of the metrics; the submission is then reviewed; and a holistic ‘best-fit’ judgement is made against the rating descriptors, based on all the available evidence.

92. The adapted method used in the pilot is summarised in Figure 7 below. It also indicates how provider-level information was considered within subject-level assessments; and vice versa.

**Figure 7: Information used in both subject and provider-level assessment**



## Outcomes

93. Assessment in the pilot resulted in one of three possible indicative ratings: Bronze, Silver, or Gold. At subject level, subjects could also receive a ‘no rating’ decision, where the panel deemed there was insufficient evidence to make a ‘best fit’ judgement.

94. The second pilot also tested options for producing subject-level statements of findings for the first time. The aim was to test how to produce feedback statements for providers, and potentially for prospective students, that could usefully supplement the subject-level ratings and could also be feasible to deliver at scale. Each statement was written by a subject panel member allocated to review the subject at stage 1, who was also tasked with keeping a note of the collective decision made by the panel.

95. We tested two approaches to subject-level feedback statements; one for each batch of subject-level assessments:

    a. **For the first batch:** panels agreed one overall rating for each subject. The statement of findings for each subject provided a brief narrative in relation to each aspect of assessment (TQ, LE and SO)

    b. **For the second batch:** panels agreed an overall rating for the subject as well as a rating for each aspect (TQ, LE and SO). The statement of findings for each subject comprised these aspect ratings, with minimal additional narrative.

96. For provider-level assessments a statement of findings included both types of information: a rating as well as a narrative explanation for each aspect.

97. The allocation process built on established principles developed in previous TEF exercises. As set out in the guidance documentation, each submission was allocated to three individual members of the relevant panel, for individual (‘stage 1’) assessment. We retained the mix of two academics and one student assessing each submission. For provider-level submissions, at least one member was from a similar type of provider. For subject-level submissions, at least one member had relevant subject expertise (defined at CAH2 level). The allocation process also divided the submissions into two batches (by specifying which of the two meetings each submission would be assessed in).

98. This allocation approach worked well for the pilot, but a number of operational issues arose and highlighted additional challenges that would be involved if scaling up to a full subject exercise:

    a. A number of panels reflected on the importance and influence of subject expertise in assessment. Matching panel expertise to subject at CAH2 level was built into the allocation algorithms, but it was clear that in some areas – particularly very specialist subjects – the ratio of subject experts to submissions placed pressure on allocation and panel members’ specialist subject expertise did not always align well with the provision being assessed (and in particular where the CAH2 subject categories included a diverse range of courses).

    b. To trial ways of managing larger volumes of decision-making, we deliberately did not plan to include separate trio discussions (see paragraph 100). Consequently, we did not

allocate submissions in a way that could enable trio discussions. If trio discussions were to be included in future – as recommended by the pilot panels – this would substantially increase the challenges involved in allocation.

c. It was found that allocation in the pilot left little flexibility to resolve issues around attendance. Given the large number of panel members, we anticipated some level of non-attendance due to illness or other unavoidable situations. We planned that chairs and deputies (who had been excluded from the initial allocation) would be able to backfill where required. However, in reality it was difficult to spread backfill work evenly (due to conflicts of interest, and to the need to maintain the 2:1 academic to student ratio of assessors). It thus became burdensome for particular individuals and had to be managed by OfS and QAA staff on a case-by-case basis. A more scalable approach would be required in future.

99. For the second subject-level pilot, panel members completed a total of 1,477 individual 'stage 1' assessments. In stage 1, ratings were made on a 5-point scale (Gold, Silver, Bronze, and the two 'borderlines'). In 30 per cent of cases, all members of the trio independently arrived at the same rating at stage 1. In the majority of cases (54 per cent), two out of three trio members arrived at the same rating. In 15 per cent of cases, each member of the trio arrived at a different stage 1 rating. This level of trio agreement was consistent across all the panels.

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

Natural Sciences, and Engineering and Technology panel report

"… there was not as much subject-specific discussion as may have been expected. […] few panel members were familiar with the distinctive structure of architecture programmes and the implication for the interpretation of metrics in this subject […] Panel members felt it would be helpful to have contextual information about subject areas, including the role of PSRB accreditations and distinctive approaches to delivery of teaching."

Social Sciences, and Natural and Built Environment panel report

"The group of three initial assessors for each case should have subject-level expertise within the discipline."

Natural Sciences, and Engineering and Technology panel report

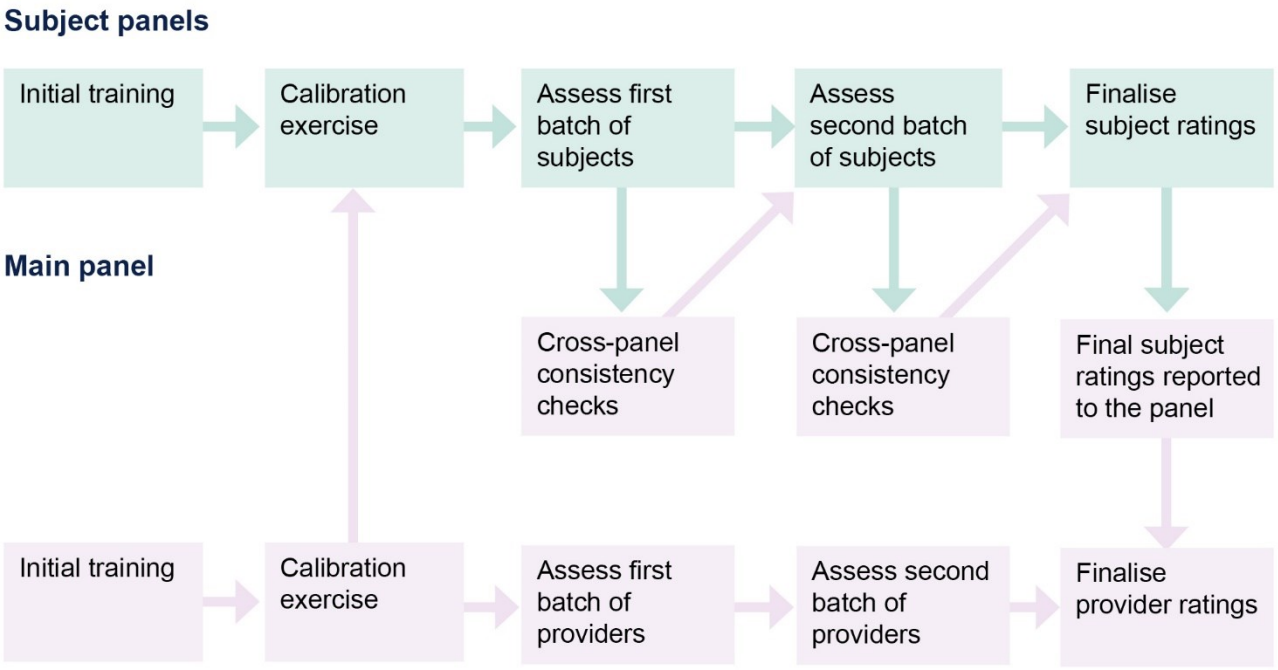
## Stage 2 and 3 processes, moderation and consistency checking

100. Following the established principles developed in previous TEF exercises, allocation was constructed to facilitate panels breaking into two groups of nine. After the individual stage 1 assessments, these groups met in stage 2 (the 'groups of nine discussions') to discuss each case and form a recommended rating. However, to test an approach that would be able to cope with the much larger volumes of assessments in a full subject TEF, we did not include the trio discussions that normally take place at the start of stage 2.

101. After the groups of nine formed their recommendations, the whole panel convened in stage 3 to consider these recommendations and make the final rating decisions.

102. For the second pilot we tested a range of processes to moderate assessments and check for internal consistency within each panel, and for consistency across the five subject panels. Figure 8 outlines the process for the main panel to check for consistency across the subject panels, and for these checks to feed back into the subject panels.

**Figure 8: Moderation activity between main panel and subject panels**




103. At stages 2 and 3 more difficult or borderline cases were discussed in greater detail, as necessary. For cases identified as more difficult during stage 2, options included:

a. Review by a fourth reviewer (and occasionally additional panel members), usually from the other group of nine. This broadly worked well although panels varied in how extensively they made use of this option, and some panel members were under pressure to review these cases in a very limited time.

b. Deferral to the whole panel for consideration or until later in the process. This enabled the whole panel to make borderline judgements in a consistent manner. However, due to timing pressures, the panels had limited opportunity to review all borderline cases together, as has been the practice in provider-level TEF.

104. To test for consistency of decisions within and between the groups of nine we also:

a. Analysed all individual ratings ahead of assessment meetings, so that the OfS could identify and report patterns of individual and group rating behaviour at the meetings

b. Allocated ‘cross check cases’, where some cases (typically with ‘borderline’ metrics, i.e. straddling Bronze and Silver, or Silver and Gold initial hypotheses) were considered by both groups of nine. Within each panel, seven to eight cross-check cases were assessed by both groups of nine. Different ratings emerged across the two groups of nine, for at least one cross-check case considered by each subject panel. This

highlighted the need for the panels to closely examine all borderline cases in each group of nine. All panels agreed that this was a useful calibration mechanism which could be expanded in future exercises.

105. The mechanisms to support consistency across the five subject panels were:

    a. Oral reports from the subject panel chairs and deputies to the main panel, where issues arising were discussed, for example approaches to not rating subjects with limited data.

    b. Main panel consideration of ratings data generated through the subject panels' first and second rounds of assessment. Comparisons were also made between the ratings arrived at for different subjects within panels, and the extent to which each panel had moved away from the metrics-based initial hypotheses through the stages of assessment.

    c. Observation across the subject panel meetings by the chair and deputies of the main panel, and OfS staff.

    d. A 'deep dive' exercise where a number of providers were examined as a whole by individual main panel members. We then compared their relative ratings of subjects with those of the subject panels.

106. These processes were considered by the main panel to be an appropriate level of checking for a pilot exercise. They revealed a broad level of consistency in the patterns of ratings across subject panels, and revealed some common issues and some differences in approach for discussion across the subject panel chairs and deputies.

107. However, some areas of apparent divergence also arose. There were limited opportunities to investigate these in detail, and limited time for guidance from the main panel to influence subsequent subject panel deliberations. For an exercise with published results, additional processes would be needed to more fully investigate issues of consistency; to determine what would constitute acceptable differences between subjects and what would require moderation; and to subsequently moderate judgements as necessary across the panels. These processes may need to include, for example, extended initial calibration activities, additional rounds of assessment meetings to allow for main panel feedback to influence subject panel assessments, and the opportunity to revisit cases that were assessed earlier in the process. Building in these additional processes for a full-scale subject exercise with many more panel members and more panels would present significant challenges and require more time.

"[…] fourth (and sometimes fifth) readers were valuable when consensus could not be reached within the trios. [..] often fourth readers were asked to focus on a specific area of the submission that the trio had found challenging in reaching consensus."

Business and Law, and Education and Social Care panel report

> "More cross-checking is required. In particular, moderating and cross-checking clear Gold, Silver and Bronze cases could be useful in determining examples of ratings at a mid-point of the rating criteria."
>
> Medical Sciences, and Nursing and Allied Health panel report

108. Panel members worked extremely diligently and all observers commented on their professionalism and constructive approach to resolving arising issues. The panels reflected carefully on the assessment processes, and the subject panel reports include a wealth of constructive feedback.

109. A clear theme emerged around timing. Panel members found that:

    a. **There was not enough time for assessment and consistency checking.** Timing for panellists was, at many points, extraordinarily tight. There was not enough time for the volume of case-by-case discussions, and there would be a real challenge in scaling up the amount of meeting time given that there would be limits to how much time people can devote to meetings outside of their other commitments. There was not enough time for full moderation either within panels or across panels. Extra iterations would be needed with the opportunity to return to cases assessed in earlier batches.

    b. **There were scheduling difficulties.** Scheduling was made difficult due to Easter falling midway through the process, which in turn meant that some meetings had to be scheduled on consecutive working days with no break. We were keen to stagger the release of information, in order to ensure that issues raised in moderating early batches of assessment could feed formatively into subsequent batches of assessment, but it was noted that this reduced the ability of panel members to manage their own workloads. Behind the scenes time was required to process documentation, for example transferring provider submissions from the OfS upload website to the panel assessment portal, and this also limited the options for scheduling meetings.

110. A second related theme concerned the staged decision-making process, which would need further refinement. With future scalability in mind, we tested a truncation of some stages of the usual TEF assessment process.

    a. **Initial discussions need to be more thorough.** In subject panels, panel members moved directly from individual assessment to 'groups of nine' discussions, skipping a separate face-to-face trio discussion that has been utilised in other TEF exercises. The subject panel reports offer numerous observations about how this was not a satisfactory process, adversely affecting the quality of decision making and making it difficult to write the statements of findings. It also in some cases made the initial discussions feel less inclusive as it placed more pressure on those who were less confident in a panel setting. In response to this feedback the main panel took an alternative approach, which did utilise face-to-face trio discussions, which was possible due to the much smaller number of cases it was assessing. The subject panels would want a full subject TEF to include some way for trios to confer prior to panel meetings. However, doing so face-to-face at panel meetings would involve several thousand trio discussions, adding a considerable amount of meeting time and significant logistical challenges. Some other mechanism would need to be developed.

b. **Group decisions would need to be more efficient.** Each panel had responsibility for the entire set of ratings it produced, which required each case to be considered by the whole group. This meant that all panel members heard the arguments and rationale developed in earlier stages for the final decision in stage 3. This repetition sometimes helped to moderate similar cases; however it was inefficient. The time required at stage 3 to process the volume of judgements required in a real exercise would further exacerbate the timing issues noted above.

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

Business and Law, and Education and Social Care panel report

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

Business and Law, and Education and Social Care panel report

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

Medical Sciences, and Nursing and Allied Health panel report

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

Social Sciences, and Natural and Built Environment panel report

"We are aware that there is a challenge of scalability […] the process may have benefited from virtual discussion between the initial assessors, or through sharing of data even if discussion were not possible. There was a sense […] that we were not quite making full use of the panel's time […] especially in stages where half of the panel is involved. [though] we recognise the need for the whole panel to be signed up to eventual ratings decisions. Changes to streamline the panel assessment process may have allowed for greater scalability and allow panel members a fuller sense of agency in the process."

Arts and Humanities panel report

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

Business and Law, and Education and Social Care panel report

"Transparency was valued and seen to support due diligence, although some panel members felt that this could have been achieved through a more thorough calibration and use of the trios. The potential for inequality of opinions by different groups (e.g. students versus academics) was raised and […] students […] did find it more difficult to comment and assess courses where they did not have specialist knowledge."

Medical Sciences, and Nursing and Allied Health panel report

"It proved difficult to co-ordinate input to SOFs [statements of findings] across the three reviewers in the assessing trio, […] It was also harder for lead reviewers to produce SOFs when they held an outlying view within the three or four assessors that had considered the submission."

Social Sciences, and Natural and Built Environment panel report

"The initial discussions between the three assessors were a useful setting to input into the process, though this does not mean we cannot contribute equally as effectively if alternative methods of deliberation are ultimately chosen for the full subject-level TEF."

Employment expert report

## Views of panel members on refinements to the framework and evidence

111. In the first pilot, two models were tested that were intended to produce ratings while reducing the burden of TEF at subject level. However, these models were found not to produce robust ratings (the model designs meant that subjects received ratings with limited or even no direct assessment).[27] In the second pilot, the comprehensive model that assessed all subjects in full was found to be an improvement on models from the year before. Panels were more confident in applying the assessment methodology and rating each subject against the criteria and rating descriptors.

112. As well as the comprehensive model, we tested a range of other refinements to the framework. We treated the pilot as an opportunity to test a wider range of changes than might ultimately be needed in a full exercise.

113. Across the panel reports, clear support emerged for a number of these refinements, including:

    a. The separation of TEF criterion 'TQ1: Student engagement' into two distinct criteria.

[27] TEF: Findings from the first subject pilot 2017-18, pages 48 to 61. Available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-findings-from-the-first-subject-pilot-2017-18/.

b. The steps taken towards strengthening student engagement with the process. Subject panels found it valuable where students had contributed to subject-level submissions, but this was patchy. The main panel found the 'student declaration' was a step in the right direction and would support extending opportunities for students to submit evidence more directly (see paragraphs 115-118 below).

c. The introduction of distinct criteria and rating descriptors at subject and provider levels. This would be further improved by clarifying the relationship between provider and subject-level ratings.

d. The inclusion of a provider summary statement as part of the subject-level evidence.

e. The addition of data on differential degree attainment, though in practice it was found to be difficult to interpret and would need to be developed further into a metric to become more useful.

"The change to the criterion to divide student engagement into learning (TQ1) and student partnership (TQ5) was positively perceived as it offered a means for distinguishing excellent practice. […]. TQ5 distinguishes those providers that take a more collegiate and responsive approach to learning and teaching."

Medical Sciences, and Nursing and Allied Health panel report

"[…] the new criteria (TQ1 and TQ5) helped the panel focus on student experience, partnership and engagement."

Arts and Humanities panel report

"The revised criteria for subject and provider-level assessment were welcomed and helped the panel reach robust judgements."

Arts and Humanities panel report

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

Business and Law, and Education and Social Care panel report

"The panel noted that the provider-level summary statements were very helpful, although they were not always well aligned with the subject submissions."

Natural Sciences, and Engineering and Technology panel report

114. The panels had mixed views on the data overall. There was a general sense of data overload, and particular pieces of data were seldom used. This was noted anecdotally throughout, and at the end of the panel assessment we polled panel members on which data – beyond the core metrics – was perceived as most and least relevant to their assessments:

    a. There was very strong support for data relating to student characteristics, and high and low absolute value markers.

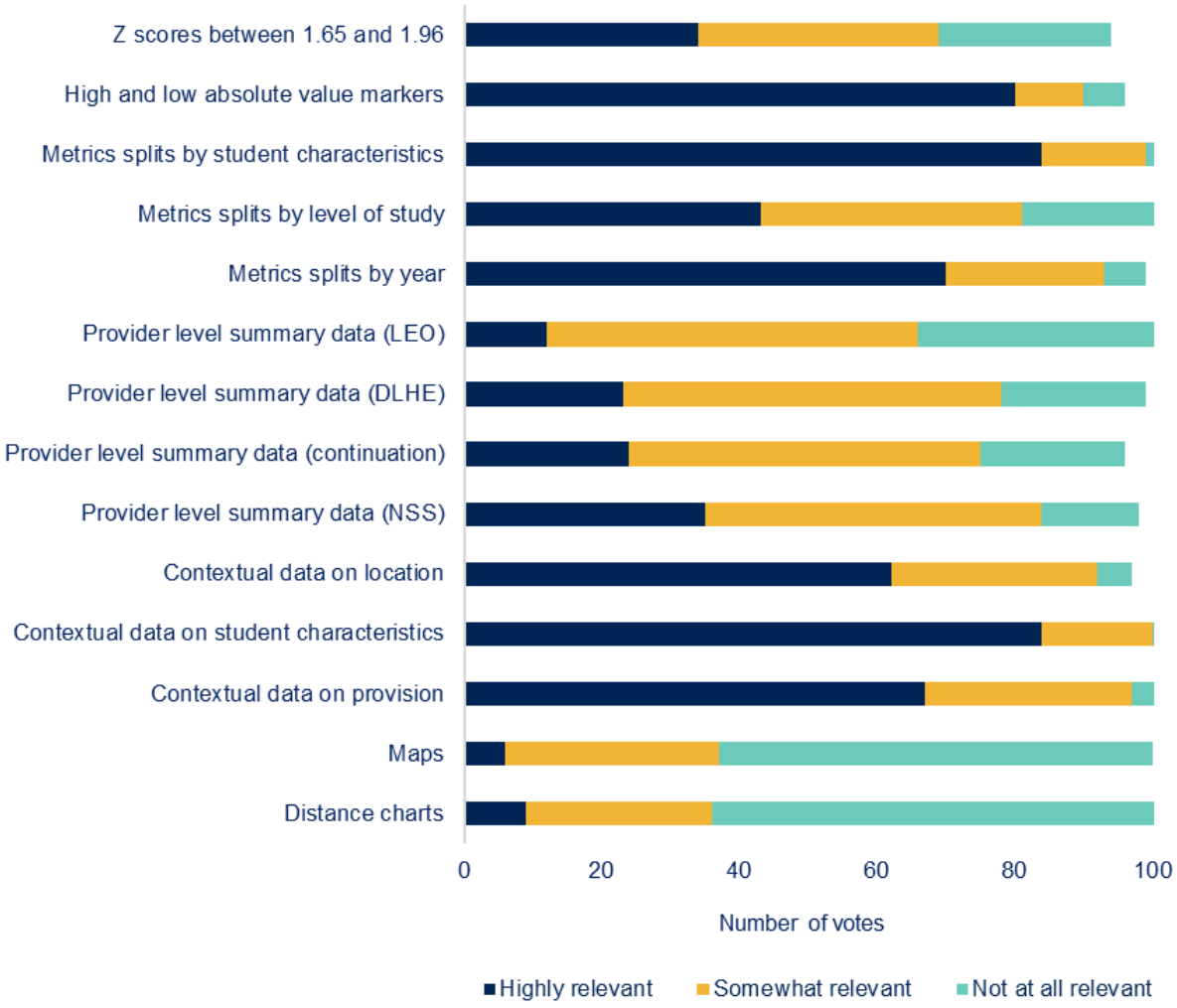
b. The employment-related maps and distance charts were found to be least useful. They were provided to allow panels to take account of geographic and mobility issues when interpreting the employment metrics. However, they were found to be complex and difficult to use. Panels would prefer geographic factors to be taken account of in the metrics themselves, for example as benchmarking factors. The employment expert report (Annex E) sets out detailed considerations for accounting for region of employment in LEO.

c. There were mixed views about the importance of access to provider-level metrics, when assessing subjects.

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

Natural Sciences, and Engineering and Technology panel report

**Figure 9: Panel members' perceptions of data relevance**




"[M]etrics and contextual information have become complex and very time-consuming to consider, especially where a dual hypothesis is required. This raises a danger of 'metric fatigue'"

Business and Law, and Education and Social Care panel report

"There was value in provider-level and subject-level contextual data, […] which contributed to in-depth discussions and changes in ratings as panel members moved through each step of the assessment process."

Medical Sciences, and Nursing and Allied Health panel report

## Student declarations

115. TEF lead student representatives from 37 providers completed a declaration indicating whether and how students were involved in submissions. These declarations were submitted directly to the OfS and were given to the pilot main panel as additional contextual information. A session was run at the first main panel meeting with student main panel members and

student deputy chairs to discuss how student declarations might be used in the assessment process. This session included:

a. A mini-calibration exercise to consider a small number of example student declarations

b. Discussion of whether any of the student declarations read by student main panel members would have affected their provider-level ratings recommendations

c. Consideration by student deputy chairs on whether the student declarations could be useful to subject panel chairs in a full exercise and how they could be used.

116. It was agreed that, for the second pilot, main panel members would try to use student declarations as contextual information for the provider-level assessment. In advance of the second main panel meeting, panel members completing provider-level assessments reviewed the submitted student declarations for all providers in their allocation. Student feedback gathered from these declarations fed into the discussion of a provider's final rating at this second meeting, and were used to corroborate evidence, or a lack of evidence, in the provider submission about student involvement in the TEF process.

117. During assessment the main panel found that the student declaration did not provide them with the range of information about student engagement they wished to know – this was captured clearly in the student findings report:

> "Panel members felt there were legitimate questions that could be asked relating to the students' involvement more broadly in the enhancement of teaching and learning. The student declaration could not accurately be used as a proxy for this, as this was not its intention when created. It was concluded therefore that, while the student declaration was successful in achieving its initial aims, it was the broader involvement of students that panel members wanted information on. It was advised that the declaration should be enhanced further as opposed to being removed from the TEF assessment process."
>
> Student findings report

118. Students concluded that the addition of an independent piece of evidence submitted by students to be reviewed in the assessment process was a positive step. However, the scope of the student declaration was limiting and restricted the main panel's understanding of student engagement in relation to TQ5: student partnership. The student deputy chair and student panel members felt that the commentary provided by students could be enhanced to provide the assessment panel a fuller picture of students' continued involvement in teaching and learning monitoring and enhancement. The main panel was supportive of the student point of view on this matter.

### Student engagement with submissions

119. From its inception, TEF and the processes within it have been designed to enable a range of forms of student engagement. Student engagement is embedded within the assessment framework and within the ratings descriptors, and so is essential for achieving higher awards.

It was no surprise that where student engagement was working well, panels quickly recognised it and judged positively its benefits.

120. As discussed in the submissions section (see paragraphs 45-72), the pilot also identified that student engagement could be patchy and that some providers and student representatives did not have the same capacity or opportunities as others to coordinate and collate input. In some cases there appeared to be evidence of poor student engagement, but in future this would not always be able to be verified or assumed without introducing further processes that might put providers and their student representatives in tension. Some subject panels also suggested that student declarations should have been made available to them as well as to the main panel. The pilot did not resolve these issues and further work would need to be done to consider the full range of implications of such issues for assessment, whilst maintaining fairness to different provider types.

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

Business and Law, and Education and Social Care panel report

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

Social Sciences, and Natural and Built Environment panel report

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

Arts and Humanities panel report

## Supplementary data

121. In the pilot we tested a developmental set of supplementary data relating to degree classifications. It was used at provider level only and applied only to providers that hold taught degree awarding powers.

122. As had been done in the first pilot, data on grade inflation was provided to show the change over time in the number and proportion of degree awarded as 1sts, 2:1s, other degree classifications and unclassified degree awards. As in the previous year, the panel recognised the wider importance of understanding and addressing grade inflation to help ensure that degrees hold their value over time. However, there was little support for this data to be

considered in TEF assessments. The main panel considered that a complex interaction of different factors could potentially explain overall changes in degree classifications, such as efforts to close attainment gaps by equality protected characteristics. This made it particularly difficult to make TEF judgements about the absolute changes in the classifications awarded by different providers.

123. Data relating to attainment gaps was provided to show the change over time in the number and proportion of degrees awarded as 1sts and 2:1s to each of the student groups shown in the splits. Panels spoke extensively of the importance of closing these gaps in outcomes and so welcomed having this information. Whilst the suite of degree classification data was cited by the main panel as not well understood, nor consistently used in provider-level assessments there was clear support for the attainment gaps data to be developed into a core metric. There was also support for the data to, if possible, be extended to also inform subject-level assessment.

"By including differential attainment data as a core metric, thus encouraging providers to address negative performance against this metric in the submission process; a diverse range of student groups will be better informed on attainment."

Student findings report

"[I]information on attainment gaps in degree outcomes (firsts and 2:1s) was available at provider level for the provider-level panel, but this information was not visible to subject panel assessors. […] It was felt that this would be useful information."

Social Sciences, and Natural and Built Environment panel report

"The provision of differential attainment data as a supplementary dataset was welcomed by all panel members, but particularly student panel members for whom this provides a key piece of information about the parity of teaching excellence at higher education providers."

Student findings report

"A small number of providers referred to their work addressing their attainment differentials [...] The panel would like to have seen addressing attainment gaps as an important focus for TEF."

Arts and Humanities panel report

"[…] it would be useful to see [the use of differential attainment data] in the TEF metrics, especially in relation to ethnicity, age, gender, POLAR/IMD and disability. The experts believe this is more relevant and useful data to assess quality than grade inflation data."

Widening participation experts report

## Use of metrics

124. The assessment process starts with panel members reviewing a provider's metrics to form an 'initial hypothesis' – an indicative rating which serves as a starting point for the panel's assessment. In step 1a, this initial hypothesis is calculated formulaically by translating the combination of positive and negative flags on individual metrics into a rating.

125. Panels in this pilot worked with a revised formula for calculating the initial hypothesis. The calculation maintained the principles previously determined by the DfE for how different metric types should be weighted, but was adjusted to reflect the greater number of metrics. In this pilot we also formalised the use of borderline ratings (i.e. Gold/Silver and Silver/Bronze) at the start of the assessment in an effort to reduce anchoring effects identified in the previous pilot, whereby panel members found it difficult to move their judgements away from the starting point of the initial hypothesis.

126. In previous years providers and panel members had to calculate the starting point for the initial hypotheses manually, but in this pilot the OfS provided the calculations from the outset in order to reduce errors and to avoid unnecessary discussion time spent confirming the calculations in panel meetings.

127. One theme of commentary on the initial hypothesis related to the weightings, and the influence this had on later stages of the decision-making process. In particular, panels questioned the half weighting of NSS metrics, the double weighting of continuation and what was seen as the overweighting of the employment metrics.

"The panel questioned whether the current system of weighting of the metrics was effective in assessing medicine and health sciences subjects. Providers could be identified as Gold initially even with poor NSS metrics, thus meaning that the initial hypothesis did not reflect the wider student experience. Greater weighting could be placed upon NSS metrics. Subject specialists observed that this could be problematic, however, with regulated courses and those including a high inclusion of practice placement time."

Medical Sciences, and Nursing and Allied Health panel report

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

Natural Sciences, and Engineering and Technology panel report

"The balance of weightings between different sets of metrics: it was suggested that continuation and employment metrics could count for relatively less than they did in the subject pilot, and teaching quality could count for more"

Social Sciences, and Natural and Built Environment panel report

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

Arts and Humanities panel report

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

Business and Law, and Education and Social Care panel report

"The panel questioned whether the continuation metric was weighted too heavily, recognising that values-based recruitment can influence continuation. Professional courses such as nursing or medicine may have higher voluntary continuation, but progression is restricted by professional fitness and competence in order to progress into future practice, which may force attrition."

Medical Sciences, and Nursing and Allied Health panel report

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

Natural Sciences, and Engineering and Technology panel report

"More student input was recommended as a limited involvement in submissions and half-weighted NSS data was not considered sufficient in highlighting the student experience. Also, perhaps students should be able to view the 'per aspect' ratings with relevant supporting information."

Medical Sciences, and Nursing and Allied Health panel report

"Employment: It was felt that the three core metrics on employment/employability gave too much weighting to this particular aspect of quality. Notably, an employer representative on the panel felt strongly that too much weighting was given to these three metrics. A reduction in the total weighting of the employment metrics need not mean less emphasis on embedding professional skills, as these should be an integral part of TQ and LE."

Social Sciences, and Natural and Built Environment panel report

"Metrics are too heavily weighted to outcomes – a greater focus on teaching excellence and student experience is needed."

Medical Sciences, and Nursing and Allied Health panel report

128. Concerns were also raised about the interaction of missing data and the starting point for the initial hypothesis. For example, it was felt that providers with missing data did not have as many opportunities to receive positive flags, thus supressing the numbers of Silver/Gold or Gold starting points. As noted in the Summary ratings and analysis section (paragraphs 161-218), the OfS analysis demonstrates that FECs and other smaller providers were disproportionately affected by missing metrics.

129. Another general concern was that the formula for the starting point for the initial hypothesis set the bar for Gold too high. With nine metrics rather than six, it was assumed that there was a higher likelihood that providers would trigger the clause that there must be no negative flags to permit a Gold starting point. Analysis of the initial hypotheses generated in the pilot (see paragraphs 161-218) suggests that this was largely mitigated by the use of borderline ratings as starting points. However, the tendency towards Silver starting points at subject level remained despite the introduction of borderline ratings.

130. In terms of how the panel operationalised the expanded basket of metrics and contextual data as a whole, there was a sense that more data could often introduce too many contradictory signals into the assessment process. With more information at hand, panels more frequently found and interrogated areas of weakness.

131. Figure 11 which is based on all providers, gives a sense of these issues, demonstrating that at subject level, 61 per cent of cases had at least one negatively flagged metric and 43 per cent of cases had a mixture of positive and negative flags (this was an increase compared to Year Two of provider-level TEF where the equivalent figures were 52 per cent and 30 per cent respectively). Panels judged that, where providers had overlooked or not responded effectively to areas of weakness, it was right that they should reflect this in their decisions. But there was a sense that the increased visibility of data suggesting poor performance may have shifted the balance towards more pessimistic assessments.

"The relatively low proportion of Gold ratings may be an artefact of the particular sample of providers and subjects in this pilot. However, an eye needs to be kept on this in any scaled-up subject TEF. The panel noted that the basket of metrics made achieving Gold in the early stages of assessment exceptionally difficult compared to Bronze or Silver/Bronze, to the extent that the panel tended to burst into spontaneous applause when a subject 'made it' on the core metrics. Given our deep conviction of the high international quality of the UK higher education system, we feel it would not be unreasonable to equalise the chances of starting as Gold or Bronze on the metrics. Panels are perfectly capable of demoting as well as promoting cases on a holistic basis, so we feel the risk is small."

Arts and Humanities panel report

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

Business and Law, and Education and Social Care panel report

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

Business and Law, and Education and Social Care panel report

"Negative performance in one metric at step 1b should not preclude a subject being awarded a Gold TEF rating. The panel felt the formulaic method for arriving at a starting point rating was too limiting in terms of the number of subjects which started at Gold."

Natural Sciences, and Engineering and Technology panel report

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

Natural Sciences, and Engineering and Technology panel report

"The main panel observed that the differences between the TEF ratings awarded for each aspect of quality to different types of provider was larger for the Student Outcomes aspect than for Teaching Quality and Learning Environment. For example, no further education colleges were rated Gold for Student Outcomes at provider level or subject level. This was the only aspect of quality where this was the case. If students at further education colleges and higher education institutions differ in their social background in ways not captured by POLAR4, this difference in ratings may have been driven by the metrics not successfully taking such factors into account."

Employment expert report

### Specialist panel roles – employer and PSRB representatives

132. Each of the 10 subject panels included up to two representatives from employers or PSRBs. Employer and PSRB representatives were trained in the assessment process, and contributed to assessment decisions, but were given more flexibility regarding case selection, and the evidence or criteria they focused on. It was generally felt that this approach made better use of representatives' expertise than simply using them as additional panel members, with a more rigid allocation and remit.

133. The main challenges that representatives with a specialist role identified with included:

    a. The time commitments required from busy professionals with less direct interest in the TEF than academic participants.

    b. The steep learning curve required to get to grips with a complex process and to develop sector-specific knowledge (for example, decoding jargon and other terminology).

    c. Ensuring representative views were consistently called upon and taken into account, in light of the more flexible, less formalised role.

134. There was a general opinion that, despite the challenges, the externality and alternative perspectives provided by employer and PSRB representatives at subject level were a helpful feature of the process, particularly as a counterpoint to the academic backgrounds of other panel members.

135. A common observation from PSRB representatives was that panel members might have benefitted from training which explained the requirements and accreditation processes of PSRBs in their subject areas. Information about accreditation for tightly regulated subjects was one of the few areas where there was a consensus that additional mandatory evidence would have been welcomed.

### Specialist panel roles – widening participation liaisons

136. Each of the 10 subject panels had one panel member appointed as the widening participation (WP) liaison. Each WP liaison was an academic or student member of their subject panel, with a full caseload. The role of the liaisons was to help ensure that WP issues were considered in all cases by reminding panel members of their responsibility. Liaisons were not expected to have expert knowledge of WP, and they did not have a remit to specifically review submissions from providers with a particular WP focus or challenge.

137. The role was generally well-received and understood within panels. Panels particularly welcomed the liaisons prompting discussion of WP issues, as they were not always considered in the current steps of assessment. However, it was felt that a more structured approach to supporting and challenging discussion of WP would be helpful, such as a list of discussion prompts. While it was not a required part of their role, some of the WP liaisons also became ‘extra readers’ on tricky cases with a WP element to them, which added somewhat to their workload.

138. It was observed that further training on the metrics would have been beneficial, in particular on the nuances within, and between, POLAR and IMD as measures of disadvantage. This would provide greater reassurance for all panel members that claims made regarding WP populations in the provider submission were supported by evidence in the metrics and contextual data.

### Specialist panel roles – interdisciplinary liaisons

139. Each subject panel had a panel member appointed as the interdisciplinary liaison. Each interdisciplinary liaison was an academic member of their subject panel, with a full caseload. The role of the liaisons was to help ensure that interdisciplinary issues were considered in all

relevant cases. Interdisciplinary provision falls broadly into two categories: joint honours programmes; and single honours programmes, such as Natural Sciences, that are 'genuinely' interdisciplinary in nature.

140. These panel members were supplied with data illustrating where students were attributed to multiple subjects. Interdisciplinary liaisons found the data to be useful but were not always clear how to interpret it. For the role to be effective, some interdisciplinary liaisons suggested that more targeted data, for example a bespoke workbook and narratives, would be required. Although this might increase confidence, it would also make the analysis time-consuming and more complex. An alternative approach would be to provide the data to enable interdisciplinary subjects to be compared with similar subjects in other institutions.

141. In many cases, there were small numbers of interdisciplinary students masked by larger metric populations. This made it difficult to discern the particular contribution of the interdisciplinary provision and to be confident in coming up with ratings which were accurate and robust.

142. There was little formal interaction between interdisciplinary liaisons in different panels. In future, consideration could be given to selecting a specific interdisciplinary liaison on each panel who deals only with those cases identified as being interdisciplinary. The focus would then be on reviewing a smaller number of cases in greater depth. In future, interdisciplinary liaisons would benefit from additional guidance and training to better understand how to make use of the data and what their role is within the panel.

143. Although there were some examples where providers devoted space to discuss interdisciplinary provision, there was, in many cases, little or no mention of it in written submissions. Particular challenges for providers included:

- lack of dedicated space to write about interdisciplinary provision
- how to address cases where there were multiple subjects with different metrics profiles
- understanding programme mappings
- the resource-intensive nature of coordinating responses
- providing adequate evidence at a more granular level.

Given the challenges in writing about interdisciplinary provision, providers might find specific guidance helpful on how to articulate the advantages of interdisciplinary provision, and the structures and support in place for those students.

144. Most interdisciplinary liaisons agreed that, in cases with 'high' interdisciplinary provision, providers could be granted an extra half or full page in their submission to discuss explicitly the implications of interdisciplinarity and how it is supported. To determine what is meant by 'high' provision, there could be an agreed threshold level above which providers would be required to write about interdisciplinary provision at either provider level or in the 'major' subject area while taking account of the metrics in the partner subjects.

## Assessment outcomes

**For the majority of submissions, panels felt confident that they were able to operationalise the process and use the criteria to reach robust judgements.** The quality of deliberations, and impact on outcomes re-confirmed the central importance of the peer review process. The greatest concerns related to handling cases with insufficient data and evidence.

**The pilot ratings revealed significant variation in performance within providers**. It was common for providers to receive the full spectrum of ratings (Gold, Silver and Bronze) across their subjects. This confirmed that within individual providers there are both areas of excellence to be celebrated and areas where they would want to improve.

**There were Gold, Silver and Bronze ratings amongst all types of providers, but the panels were concerned about the profile of outcomes for different types of providers.** Panels noted that distinct patterns of ratings emerged for different types of providers. They reflected on the extent to which they were observing sampling effects, issues arising from differences in the data, the impact of funding disparities, or other differences. Subsequent OfS analysis confirmed that subjects in FECs and providers with fewer than 10 subjects in the pilot were likely to receive lower ratings, even when accounting for initial hypothesis performance.

**The relationship between provider-level and subject-level assessment was improved but needs further consideration.** The two models piloted in 2017-18 prescribed the relationship between provider-level and subject-level assessments. In one, subject-level assessment was based on whether a subject performs differently to the provider as a whole. In the other, subject ratings fed into the assessment of the college or university as a whole. In the 2018-19 pilot the assessments at both provider and subject level were more independent of each other. The panels supported this approach in principle, and welcomed the distinct criteria and rating descriptors for each level of assessment. However, in practice the model produced some provider and subject-level outcomes that might not be explainable to prospective students. Further work on the model would be needed to address this.

**More granular ratings and feedback statements were tested.** The panels and providers welcomed this, but further work would be needed to develop a consistent approach to producing useful feedback at scale. The pilot tested two ways in which panels could give more granular feedback beyond the Gold, Silver or Bronze ratings. Panels tested judging the three individual ‘aspects’ of teaching excellence (teaching quality, learning environment, and student outcomes and learning gain), alongside producing a single overall rating. They found this helped to mitigate the high variation between the overall ratings of different subjects. They were also keen to offer written feedback and tested this for the first time at subject level. Providers generally found the additional granularity and feedback helpful, and would welcome more detailed written feedback in future. However, the panels found it challenging to ensure feedback was accurate and consistent across the many authors involved, and there were concerns about scaling up this process. Further development would be needed to find an approach that produces feedback on a larger scale that is both consistent and useful.

## Panel views on the outcomes

145. As noted, panels developed collective confidence over time but it was clear that it took some individual members longer to be fully engaged in discussions. As noted elsewhere, more thorough preparations and support would help mitigate this. Whilst many members were clearly comfortable with the challenging deliberations required, and with interrogating their own and others' opinions, observers noted that panels developed various group dynamics throughout the course of assessment.

146. In particular, it was observed that student perspectives could sometimes differ from academic views and careful chairing was required to navigate these discussions. OfS analysis of rating behaviour indicated that at the individual assessment stage, students' ratings were on average marginally higher than academics' ratings. Observers also noted how students usefully challenged preconceived ideas and conventions. One example of this was students focusing attention on the influence (or otherwise) subjects might have over institution-wide learning resources. Another was their influencing panel interpretations of 'positive outcomes for all' and differential attainment data.

147. Ultimately, all decisions were reached through discussion, compromise and group consensus. However, it was interesting to note that often the compromise to differing views led to the decision to provide additional granularity to the rating. This was either done through the ratings for the three aspects or narrative comments in the statement of findings. Panels were keen that more nuanced judgements should be conveyed back to providers than through an overall rating alone.

148. There was consensus across all panels that the majority of their judgements were robustly made according to the process and criteria. There was less confidence where there was missing data or smaller student cohorts. Analysis shows that certain types of provider (typically FECs or providers with fewer than 10 subjects) were more affected by such issues. Panels reflected on the distributions of ratings between different types of providers and the extent to which they were observing sampling effects, issues arising from differences in the data, or other differences.

"The main panel members felt the staged, in-depth discussions led to more robust decisions being made."

Main panel chair's report

"The whole panel was in consensus that robust judgements were made. Each case was well considered and outcomes embraced a range of thoughts and views. Panel members valued each other's diverse perspectives, and especially student members, who the panel identified as 'excellent', offering critical consideration and a different viewpoint."

Medical Sciences, and Nursing and Allied Health panel report

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

Natural Sciences, and Engineering and Technology panel report

"The panel struggled to have confidence in judgements made for cases at the lower threshold for inclusion in subject-level TEF (student populations of approximately 20)."

Arts and Humanities panel report

"As the main panel met to reflect on the ratings awarded at the end of the second subject-level pilot, we considered that larger providers and typically universities were more likely to achieve a higher final rating than smaller providers within this sample. There may be legitimate reasons for this observation, such as lack of regionally benchmarked LEO data […]."

Main panel chair's report

### Relationship between provider-level and subject-level ratings

149. In the pilot, assessments were conducted at each level separately, but we tested the following linkages between the two, with the aim achieving an appropriate coherence between them:

    a. A two-page provider summary statement, which was made available to subject panels in order to situate the subject in the institutional context

    b. Providing a summary of provider-level data in subject-level assessment

    c. Providing data on the number of students taught in each subject in provider-level assessment

    d. Splitting provider-level metrics by subject as well as year and student characteristics.

150. Importantly, at the final stage of assessment, the main panel considered the profile of subject ratings for each provider, and whether that should in any way influence the provider-level rating. In general, the main panel considered that differences between the provider rating and the subject-level profile were permissible, in part because the criteria and the ratings descriptors were articulated differently at each level.

151. In some cases there was an apparent mismatch between the provider rating and subject-level profile. This included cases where a provider's overall rating was different from the rating most often received by its subjects, and cases where a provider was awarded a rating at one end of the spectrum, but had a number of subjects rated at the other end. The main panel was generally comfortable with these variations, and took the view that a provider's overall rating should not be influenced in a formulaic way by its subject ratings.

152. However, a small number of cases arose where the subject-level ratings and the rating for the provider lacked coherence. In particular, two providers received provider-level ratings that were higher than all of their subject-level ratings. Each of the panels involved confirmed

that the ratings were correct according to the evidence and criteria at each level. However, the main panel recognised that if a future subject-level TEF led to published ratings of this kind (for example, a Gold-rated provider which had only Silver-rated subjects), they might be inexplicable to prospective students.

153. Overall, while the separation of provider and subject-level assessment in this pilot was supported, it produced some outcomes that might not be explainable. Further work would be needed to develop a model that maintains distinct evidence and criteria at each level of assessment, avoids a formulaic link between them, and which also ensures the outcomes it produces are explainable.

## Statements of findings

154. In the pilot we explored how to generate useful feedback to providers, in a way that could be scalable in a full subject-level TEF. We trialled three types of statements of findings (SOF):

    a. **Type 1 (subject-level only):** Panels agreed one overall rating for each subject. The statements of finding for each subject provided a brief narrative in relation to each aspect of assessment (TQ, LE and SO). This SOF did not include a separate rating for each aspect.

    b. **Type 2 (subject-level only):** Panels agreed an overall rating for the subject as well as a rating for each aspect (TQ, LE and SO). The SOFs for each subject comprised these aspect ratings, with minimal additional narrative.

    c. **Type 3 (provider-level only):** This type of SOF combined both of the above content. The SOF featured separate ratings as well as narrative feedback for each aspect. It also included an overall best-fit holistic rating.

155. We anticipated from the outset that operationalising SOFs would be challenging. Although panels were keen to provide feedback, they also raised concerns early on about the workload, scalability and consistency of generating narrative SOFs at subject level. These concerns proved to be the case and a number of lessons were learned:

    a. In the pilot, a SOF author for each case was identified and they were typically the ‘lead reviewer’ who had been allocated the case for initial review on the basis of their subject expertise aligning with the submission. However, lead reviewers found it hard to capture the required level of notes for the SOF whilst being required to fully engage in deliberations.

    b. Distributing the writing of narrative SOFs amongst multiple panel members also significantly increased their variability in terms of tone and quality. Not all panellists had the same previous experience of writing formal and technical documentation. This would become substantially more challenging in a full-scale subject-level exercise, with three to four times as many panel members.

    c. These issues may be mitigated by having fewer SOF authors, and consideration could be given to a dedicated panel role for this purpose.

156. SOFs were released to providers in July and feedback was gathered by IFF through a series of further surveys. Overall, the IFF report found that:

    a. TEF main contacts were generally more satisfied than the academic leads (who wrote or contributed to subject-level submissions) with how findings were presented across all types of SOFs. 50 per cent of TEF leads said they were satisfied or highly satisfied, versus 33 per cent of academic leads who were satisfied or highly satisfied.

    b. Half of TEF main contacts were satisfied that the findings were a fair representation of their provision. TEF main contacts described the representations of findings as fair, and referred to the SOFs aligning with internal evaluations. Where per-aspect judgements were made (types 1 and 3) these judgements were valued. Other TEF main contacts reported dissatisfaction with the representations of findings in the SOFs, questioning the consistency of panel decisions across subjects and pointing to an insufficient level of detail in the SOFs.

    c. Opinions varied from survey respondents (TEF main contacts and academic contributors) on the different types of SOFs. TEF main contacts were the most satisfied with type 3 SOFs and this was their preferred SOF type. Type 2 SOFs were the least popular amongst respondents, with lower levels of satisfaction reported by TEF main contacts and academic contributors.

157. Survey respondents were also asked about the range of intended actions they might take as a result of having seen their SOFs. Whilst some survey respondents reported that the SOFs did not have enough detail for them to be able to act upon the findings, nearly all TEF main contacts (95 per cent) and the majority of TEF academic leads (86 per cent) reported that they had learnt something about how they would approach a future TEF exercise from participating in the subject-level pilot. Learning points included what could be included in a submission and the required resource to complete a submission.

"We will review aspects which were commented on to improve teaching, learning and assessment. We'll also produce reports for senior leaders and academic staff."

TEF main contact, Further Education College
TEF subject-level pilot evaluation – Provider perspectives

"We won't take action, as there is insufficient detail to know what actions, within our control, would enhance any future audit exercise."

Academic contributor, Specialist University
TEF subject-level pilot evaluation – Provider perspectives

"We learnt that it needs to be led by a small senior team. It is important to allow time for the guidelines and metrics to be understood. We also realised an in-house data and policy expert was central for the academic contributors to digest the elaborate TEF guideline and complex metric dataset."

TEF main contact, University
TEF subject-level pilot evaluation – Provider perspectives

"We need to: understand and focus on metrics through the academic year and focus our annual monitoring and enhancement processes around subject TEF. We need to focus on split metrics and provide strong alternative evidence for and directly address below benchmark metrics."

TEF main contact, Specialist university
TEF subject-level pilot evaluation – Provider perspectives

158. Some providers, on receipt of their awards, raised further queries directly with the OfS about the coherence and consistency of the overall package of ratings they received. One common type of query was to directly compare the metrics profiles for two or more subjects, for example highlighting where two subjects with seemingly similar metrics had received different awards. Whilst it is essential to remember that in all these cases different subject submissions would have influenced the judgements, these queries are of interest as they highlight the importance of an effective moderation process in a full-scale exercise, as providers may be more able than the panels to identify inconsistencies across subject ratings. Each provider would inevitably scrutinise its own set of results with a detailed understanding of the similarities and differences in the metrics and submissions across its subjects.

159. Another set of provider queries sought to understand whether specific arguments had been deciding factors in judgements. A typical example was where a provider had argued that a negative flag should be discounted and whether the panel had agreed with this argument. These queries perhaps indicated a tension between the holistic nature of judgements made by panels and a desire by providers to understand how specific details influenced the judgements.

160. For the pilot, a range of baseline checks were put in place prior to the release of SOFs, but ultimately a very strict quality control process would be required for a full exercise. During the pilot, the OfS only received one query that, upon investigation, identified that an error had been made in the drafting of statements of findings.

## Summary ratings and analysis

**The metrics and ratings at subject level remained more heavily concentrated at Silver, than at provider level.** The strong tendency towards a Silver initial hypothesis at subject level was reported in the first pilot.[28] We sought to mitigate this by narrowing the range of flag combinations that produce a Silver initial hypothesis, and introducing two 'borderline' categories (Gold/Silver and Silver/Bronze). However, the tendency towards Silver metrics remained. Also, the pilot ratings at subject level were less likely to move by a whole grade away from the initial hypothesis than at provider level.

**Further analysis showed that the majority of subjects across the sector do not have large enough student cohorts to robustly generate flags or consistently generate an initial hypothesis.** The analysis showed that under the current method, a subject would need to have several hundred students in order to robustly generate a flag. Only a minority of subjects have large enough cohorts for this. While panels had a general awareness that metrics based on smaller cohorts had limitations, they conducted their assessments without knowledge of the extent of this limitation, as this analysis was only carried out after the assessments had concluded. It would be vitally important to consider carefully a range of technical and policy responses to this issue.

**At subject level, the pilot ratings differed by type of provider.** After taking into account differences in their metrics and other factors, a statistically significant difference in the distribution of ratings remained (though it should be noted that student characteristics were found not to have a significant effect when other variables such as provider type and initial hypothesis were taken into account). Subjects at larger multi-faculty providers more frequently received Gold ratings, and were more likely to receive a higher rating than their initial hypothesis. Subjects at FECs and providers with a narrower range of subjects more frequently received Bronze ratings, and were more likely to receive ratings that were lower than their initial hypothesis. OfS analysis shows that these patterns remained after taking into account differences in the metrics and other measurable factors. There could have been other factors, as well as the general composition of the pilot sample, contributing to these patterns. Further work would be required to explore such differences.

**Panel decisions about 'no ratings' indicated multiple sources of uncertainty in the data at subject level.** Panels judged, on a case-by-case basis, that seven per cent of subjects contained insufficient evidence to receive a rating. This was more likely where metrics were unreportable, and where cohorts were smaller, but these two factors alone do not explain the 'no rating' decisions. Where panels did give ratings for subjects with unreportable metrics or smaller cohorts, these were more likely to have a Silver initial hypothesis and be rated Silver than subjects with more complete evidence.

[28] Findings from the first subject pilot are available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-findings-from-the-first-subject-pilot-2017-18/.

**Analysis of the data indicated similar patterns of ratings across the subject panels but with some areas of potential inconsistency.** Subjects assessed by one of the five subject panels were likely to receive lower ratings. Analysis took into account the metrics and other measurable factors, but not the strength of submissions. There were also significant differences in the proportion of 'no ratings' awarded by a number of the panels, which were not explained by differences in the data.

161. This section provides summary information about the ratings that were generated in the second subject pilot (2018-19), and other analysis of the data. In the pilot, the main panel undertook an overall provider-level assessment for 45 providers. Subject panels assessed a total of 630 subjects. The number of subjects per provider varied from one (single-subject providers) to 32, with a mean of 14 subjects per provider.

162. The data and analysis in this section was produced by the OfS after the conclusion of panel assessment meetings. It comprises:

    a. Charts showing the distribution of final ratings and the movement of these ratings from the initial hypothesis[29]

    b. Analysis of factors that might have influenced the outcomes[30]

    c. Analysis of how flags are generated for metrics based on different cohort sizes.[31]

163. The rating information shown in this report has been anonymised, de-identified and aggregated in accordance with the terms and conditions of participation agreed with participating providers in advance of the pilot.

164. Readers should exercise caution in interpreting these results. Any correlation between a certain variable and the likelihood of a higher or lower rating does not necessarily demonstrate a causal link between the two. It is not possible to infer that the distributions shown here would be replicated in a full-scale assessment due to the small and selective samples included within the pilot.

165. Readers should also exercise caution by noting that the underpinning statistical elements of the methodology have been reviewed by the Office for National Statistics (ONS) as part of the independent review of the TEF. We expect to revise the metrics methodology in the future in response to:

    - recommendations made as a result of the ONS review
    - feedback we have received in response to the pilots
    - the wider release of subject-level data

[29] The full range of charts based on the pilot ratings are included in Annex F: TEF subject-level pilot outcomes analysis.

[30] This analysis is included in Annex G: Logistic regression models.

[31] This analysis is included in Annex H: Type I and II errors in TEF.

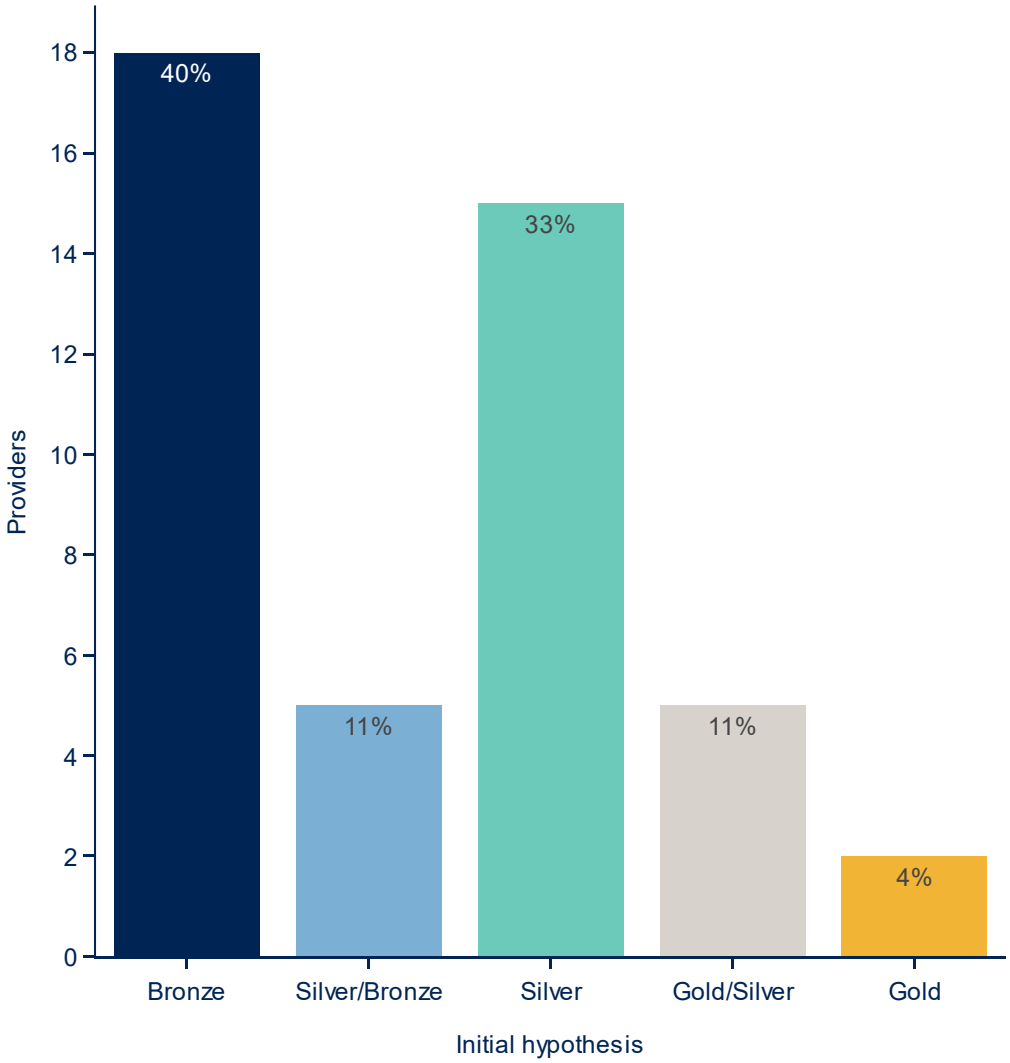
- wider changes to the data landscape (such as the transition to Graduate Outcomes data and development of LEO data).

The metrics methodology and the definitions used in the pilot and in this analysis are therefore subject to change for future TEF exercises.

## Metrics-based initial hypotheses

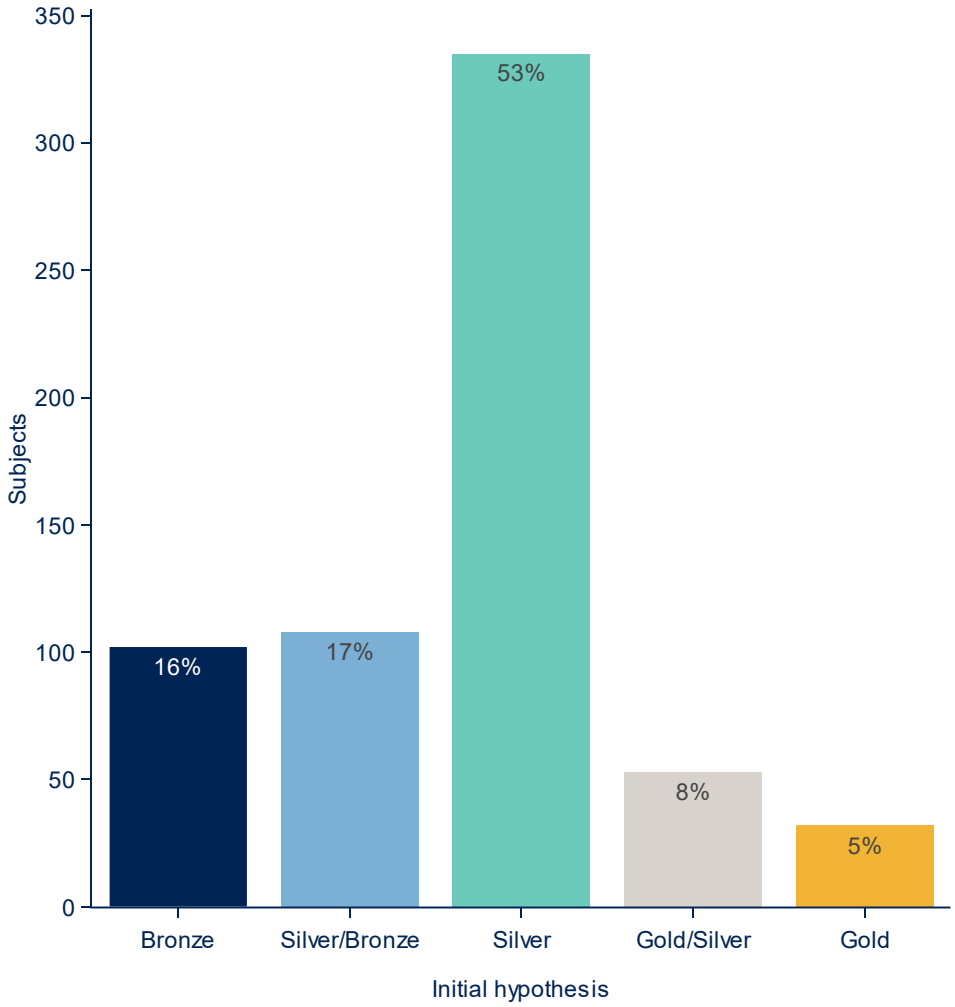
166. The distribution of initial hypotheses for the sample of providers involved in the pilot is shown in Chart 1 below. Only two providers had a Gold initial hypothesis, while 18 providers had a Bronze initial hypothesis.

### Chart 1: Provider-level initial hypothesis (without absolute values) by number of providers




167. Comparing this with the distribution of initial hypotheses for all providers in the sector shows that the metrics for the sample of providers selected for the pilot were not entirely representative of the sector as a whole. Based on pilot data across the sector as a whole: 37 per cent started as Silver; 25 per cent started as Bronze; and 7 per cent started as Gold (see Figure 10 below).

## Chart 2: Subject-level initial hypotheses (without absolute values)




168. The pilot assessed 630 subjects. The distribution of subject-level initial hypotheses in the pilot was more in line with the sector as a whole. Just over half of subjects in the pilot had a Silver initial hypothesis (53 per cent), with only 86 subjects (13 per cent of the total) having a higher starting point than this.

169. In the first pilot we reported on the skew towards a Silver initial hypothesis at subject level. Expanding the basket of metrics for the second pilot involved a risk of an increased skew towards Silver. In addition, the panels raised a concern that, with an expanded basket, it could become more difficult to achieve a Gold initial hypothesis by generating a single negative flag.

170. We sought to mitigate this issue by reducing the range of flag combinations that produce a Silver initial hypothesis and by introducing borderline categories.

171. Figure 11 shows the range of flag combinations across the sector at subject level, and the associated initial hypotheses. It shows that the vast majority of subjects that attained enough positive flags for Gold, but had a single negative flag, would have fallen into Gold/Silver borderline category.

**Figure 10: Number of flagged metrics for all providers in the sector[32]**

| Total value of negative flagged metrics | Total value of positive flagged metrics | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **0.5** | **1** | **1.5** | **2** | **2.5** | **3** | **3.5** | **4** | **4.5** | **5** | **5.5** | **6** | **6.5** | **7** | **7.5** |
| **0** | 13 | 4 | 13 | 4 | 24 | 8 | 15 | 8 | 6 | 4 | 2 | 2 | 2 | 2 | | |
| **0.5** | 4 | 1 | 4 | | 7 | | 5 | 2 | 1 | 1 | | | | | | |
| **1** | 10 | 7 | 7 | 2 | 20 | 9 | 10 | 11 | 8 | 6 | 1 | 1 | | | | |
| **1.5** | 9 | | 5 | 1 | 6 | 1 | 2 | | 1 | | | | | | | |
| **2** | 16 | 3 | 10 | 3 | 5 | 1 | 2 | | | | | | | | | |
| **2.5** | 13 | 1 | 6 | 1 | 5 | 1 | 3 | | | | | | | | | |
| **3** | 10 | 2 | 5 | 2 | 1 | 1 | 2 | | 1 | | | | | | | |
| **3.5** | 10 | 1 | 6 | | 3 | | 1 | | | | | | | | | |
| **4** | 3 | 2 | 2 | | | | | | | | | | | | | |
| **4.5** | 5 | | 1 | 1 | 2 | | | | | | | | | | | |
| **5** | 1 | | | | | | | | | | | | | | | |
| **5.5** | 6 | | 2 | | | | | | | | | | | | | |
| **6** | | | 1 | | | | | | | | | | | | | |
| **6.5** | 2 | | | | | | | | | | | | | | | |
| **7** | | | | | | | | | | | | | | | | |
| **7.5** | 1 | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | Number | 97 | 66 | 145 | 59 | 26 |
| | | | | | | | | | | | Total | | | | | 393 |
| | | | | | | | | | | | Percentage | 25% | 17% | 37% | 15% | 7% |

[32] This data covers all providers in the sector that are within scope for TEF assessment, with provider-level metrics that meet the minimum thresholds for assessment, and is based on the initial hypothesis at step 1a, excluding absolute value markers that count as flags.

**Figure 11: Number of flagged metrics for all subjects in the sector[33]**

| Total value of negative flagged metrics | Total value of positive flagged metrics | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **0.5** | **1** | **1.5** | **2** | **2.5** | **3** | **3.5** | **4** | **4.5** | **5** | **5.5** | **6** | **6.5** | **7** | **7.5** |
| **0** | 420 | 79 | 190 | 73 | 258 | 86 | 110 | 54 | 53 | 38 | 15 | 17 | 12 | 8 | 1 | 2 |
| **0.5** | 91 | 9 | 51 | 5 | 50 | 7 | 21 | 6 | 7 | 2 | 3 | | | | | |
| **1** | 219 | 52 | 107 | 52 | 120 | 34 | 50 | 17 | 21 | 10 | 4 | 3 | 1 | 2 | | |
| **1.5** | 77 | 3 | 36 | | 30 | 3 | 14 | | 1 | 2 | | | | | | |
| **2** | 180 | 29 | 55 | 18 | 57 | 16 | 12 | 7 | 5 | 1 | | 1 | | | | |
| **2.5** | 60 | 3 | 22 | 6 | 20 | | 4 | | | | | | | | | |
| **3** | 87 | 11 | 25 | 11 | 13 | 5 | 4 | | | | | | | | | |
| **3.5** | 47 | 3 | 16 | 1 | 9 | | 1 | | | | | | | | | |
| **4** | 32 | 4 | 13 | 5 | 3 | 2 | 1 | | | | | | | | | |
| **4.5** | 16 | | 1 | | 4 | | | | | | | | | | | |
| **5** | 13 | | 3 | | | | | | | | | | | | | |
| **5.5** | 5 | 1 | 1 | | | | | | | | | | | | | |
| **6** | 3 | | 1 | | | | | | | | | | | | | |
| **6.5** | | | | | | | | | | | | | | | | |
| **7** | | | | | | | | | | | | | | | | |
| **7.5** | | | | | | | | | | | | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Number | 448 | 509 | 1913 | 293 | 200 |
| Total | | | | | 3363 |
| Percentage | 13% | 15% | 57% | 9% | 6% |

[33] This data covers all subjects in the sector that are within scope for TEF assessment, with subject-level metrics that meet the minimum thresholds for assessment, and is based on the initial hypothesis at step 1a, excluding absolute value markers that count as flags.

**Figures 10 and 11 key**: Formula for calculating the step 1a initial hypothesis

| Starting point | Value of core metric flags |
|---|---|
| **Gold** | The value of positive flags is at least 3.5, and there are no negative flags. |
| **Gold/Silver** | The value of flags is within 1.0 of a Gold starting point. |
| **Silver** | All other combinations of flag values. |
| **Silver/Bronze** | The value of flags is within 1.0 of a Bronze starting point; or the value of negative flags is at least 2.5, but the value of positive flags is greater than the value of negative flags. |
| **Bronze** | The value of negative flags is at least 2.5, and the value of positive flags is less than the value of negative flags. |

172. However, Figure 10 and Figure 11 also show that, compared to metrics at provider level, a relatively small proportion of subjects attained enough positive flags for a Gold or Gold/Silver initial hypothesis. More generally, the metrics at subject level remained more skewed towards Silver than at provider level. Based on the pilot data, 37 per cent of providers across the entire sector started on a Silver initial hypothesis at provider level. This increased to 57 per cent of cases starting at Silver at subject level. This became even more significant if any of the nine core metrics were unreportable, as is more often the case at subject level. In cases where a single data source was unavailable, 70 per cent of subjects had an initial hypothesis of Silver.

## Cohort sizes required to robustly generate flags

173. Following the issues raised in the pilot relating to smaller cohort sizes at subject level, the OfS conducted further analysis of the relationship between cohort size and how flags are generated. We have included this analysis at Annex H due to its implications for the pilot findings as a whole.

174. The analysis looked specifically at how far the likelihood of differences in performance being flagged depends on the cohort size, in the current method of generating flags. To understand this, we created a model which tests artificial adjustments to the performance of a subject for a given metric and shows the likelihood that this adjustment in performance results in a flag being generated. The methodology, and the findings of the analysis, are set out in full in Annex H.

175. The analysis in Annex H was repeated for three different metrics. Across all three metrics it shows a small proportion of ‘false positives’ were generated (that is, where a flag is erroneously generated), regardless of cohort size. This is because flags were only generated where the difference from benchmark is statistically significant, with a z-score of at least 1.96.

176. However, the analysis shows a high proportion of 'false negatives' were generated, and that this was dependent on cohort size. For example, where the 'Teaching on my course' metric at subject level was based on over 1,000 students, a six percentage point difference between the provider's actual satisfaction rate and its benchmark correctly generated a flag 100 per cent of the time. When the cohort size decreased, the likelihood of false negatives dramatically increased. If this metric was based on 50-99 students, the same six percentage point difference would result in a false negative 74 per cent of the time.

177. The extent of false negatives varied slightly for each of the three metrics we analysed, but in general, using the current flagging method, cohort sizes of several hundred students might be needed to avoid high levels of false negatives, so that the metrics evidence used in the assessment (including the initial hypothesis) would be generated in a consistent way across subjects. During their assessments, the pilot panels were generally aware of, and sought to account for, limitations in the data at subject level and it is important to note that panels make holistic judgements based on both the metrics and submissions. The panels recommended in future there should be a much higher cohort size threshold than the 20 students used in the pilot. However, they were unaware during their assessments of the extent of the inconsistencies in flagging highlighted by this analysis, which was carried out after the assessments.

178. Because applying the thresholds implied by this analysis would substantially decrease the numbers of subjects assessed and students covered, the OfS is carefully exploring technical options for improving the application of metrics to smaller cohorts. Table 10 below shows the proportions of students and subjects within providers that would be excluded at different cohort thresholds.

### Table 10: Proportions of students and subjects excluded against cohort size thresholds

| Cohort size threshold | Proportion of subjects excluded (%) | Proportion of students excluded (%) |
|---|---|---|
| 20 | 23.3 | 1.2 |
| 50 | 40.6 | 3.5 |
| 100 | 54.0 | 7.7 |
| 250 | 72.4 | 21.4 |
| 500 | 88.4 | 47.1 |
| 1000 | 97.1 | 75.8 |

179. Different types of providers would be affected differentially by cohort size thresholds. For example, with a threshold of 500, further education colleges would have 99.9 per cent of subjects and 99.1 per cent of students excluded from subject-level assessment.

180. The impact on coverage also varies across subject categories. 'Nursing and midwifery' would have the lowest level of exclusion: a threshold of 500 would mean 41 per cent of subjects and 4.9 per cent of students would be excluded. The same threshold would result in several subjects being completely excluded from assessment, based on the current size of provision across the sector: Pharmacology, toxicology and pharmacy; Physics and astronomy; Chemistry; Materials and Technology; and Philosophy and Religious Studies.

181. A lower threshold of 100 students would not result in any subjects being entirely excluded from assessment. However, our analysis suggests that using the current flagging method, this threshold involves a high proportion of false negatives.

182. Table 11 below shows the proportion of subjects and students that would be excluded from subject-level assessment if minimum thresholds of 100 or 500 students were adopted.

### Table 11: Proportion of total provider offering subject excluded and total students in subject excluded by each subject category

| Subject category | Proportion of total sector provision excluded using a threshold of 100 | | Proportion of total sector provision excluded using a threshold of 500 | |
|---|---|---|---|---|
| | Subjects (%) | Students (%) | Subjects (%) | Students (%) |
| Pharmacology, Toxicology and Pharmacy | 48.9 | 10.2 | 100.0 | 100.0 |
| Physics and Astronomy | 49.2 | 5.0 | 100.0 | 100.0 |
| Chemistry | 47.9 | 9.5 | 100.0 | 100.0 |
| Materials and Technology | 78.1 | 41.4 | 100.0 | 100.0 |
| Philosophy and Religious Studies | 65.4 | 32.4 | 100.0 | 100.0 |
| Combined and General Studies | 83.3 | 6.1 | 97.9 | 14.4 |
| Politics | 65.2 | 14.1 | 97.8 | 88.3 |
| Agriculture, Food and Related Studies | 69.8 | 17.9 | 97.7 | 79.2 |
| General, Applied and Forensic Sciences | 67.2 | 20.0 | 96.9 | 74.7 |
| Architecture, Building and Planning | 73.6 | 20.7 | 96.3 | 78.3 |
| Veterinary Sciences | 62.5 | 19.3 | 95.8 | 75.6 |

| Subject category | Proportion of total sector provision excluded using a threshold of 100 | | Proportion of total sector provision excluded using a threshold of 500 | |
|---|---|---|---|---|
| | Subjects (%) | Students (%) | Subjects (%) | Students (%) |
| Health and Social Care | 62.7 | 16.9 | 95.3 | 64.7 |
| Allied Health | 61.0 | 16.1 | 95.2 | 76.5 |
| Performing Arts | 55.4 | 11.1 | 94.4 | 70.0 |
| Economics | 39.0 | 8.6 | 92.2 | 78.7 |
| Mathematical Sciences | 37.3 | 3.9 | 91.6 | 66.4 |
| Computing | 63.2 | 9.4 | 91.3 | 55.5 |
| Sport and Exercise Sciences | 62.6 | 8.4 | 90.7 | 52.7 |
| Medical Sciences | 41.5 | 9.5 | 90.2 | 69.0 |
| Media, Journalism and Communications | 54.9 | 12.2 | 90.2 | 61.8 |
| Sociology, Social Policy and Anthropology | 46.5 | 6.5 | 89.7 | 62.3 |
| English Studies | 28.2 | 3.5 | 88.9 | 66.4 |
| Languages and Area Studies | 44.4 | 6.1 | 88.9 | 63.3 |
| Biosciences | 48.6 | 4.6 | 87.0 | 57.8 |
| Geography, Earth and Environmental Studies | 39.0 | 4.6 | 86.6 | 56.5 |
| Education and Teaching | 66.0 | 10.7 | 85.6 | 32.6 |
| Psychology | 32.0 | 4.1 | 85.2 | 54.8 |
| Engineering | 59.3 | 9.5 | 84.4 | 32.2 |
| Creative Arts and Design | 58.2 | 6.3 | 82.4 | 25.1 |
| Law | 27.1 | 2.8 | 82.2 | 54.5 |
| History and Archaeology | 29.6 | 3.9 | 78.6 | 39.5 |
| Business and Management | 51.9 | 4.2 | 74.4 | 18.9 |
| Medicine and Dentistry | 28.2 | 1.9 | 56.4 | 16.5 |

| Subject category | Proportion of total sector provision excluded using a threshold of 100 | | Proportion of total sector provision excluded using a threshold of 500 | |
|---|---|---|---|---|
| | Subjects (%) | Students (%) | Subjects (%) | Students (%) |
| Nursing and Midwifery | 30.1 | 0.8 | 41.0 | 4.9 |

183. Overall, the analysis shows that, using the current TEF methodology, a subject would need to cover several hundred students in order for the metrics flags to consistently inform the assessments. Further work would be needed to consider alternative approaches and methods, and explore how metrics might in future inform assessments at subject level.

## Absolute values

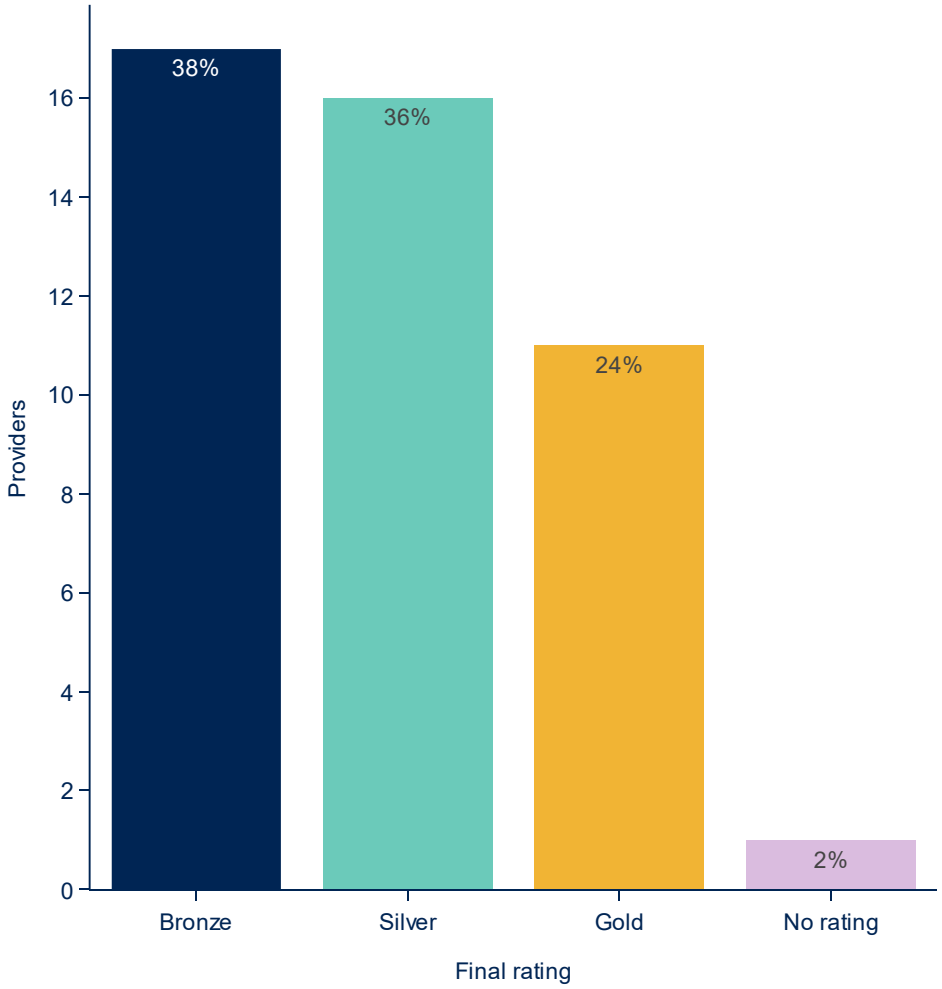
184. Very high and very low absolute values were not counted in the starting point of the initial hypothesis, but were expected to be manually accounted for by the panels when reviewing the metrics and making an overall judgement about the initial hypothesis. At provider level, only one provider's initial hypotheses changed as a direct result of adding in absolute value markers. At subject level, the impact was bigger, likely because fewer metrics received a flag or were not correctly flagged because of false negatives in the current flagging approach (see Annex H). Therefore, absolute value markers were more often taken as an indicator of performance instead. The initial hypothesis for 101 subjects (16 per cent) changed directly as a result of taking into account absolute values. The majority of these (59 per cent) were subjects that started as Silver. Rating movement as a result of absolute values seemed to be most prevalent at the boundary between Gold and Silver (this pattern is illustrated in Chart 5 and Chart 6 below).

185. Analysis of the impact of absolute value markers on the initial hypothesis alone may underestimate their impact on the process. In some cases, absolute value markers may have moved the initial hypothesis closer to a borderline without substantively changing the rating, or may have influenced the panel's holistic judgement. However, OfS regression analysis tested the impact of absolute values beyond their impact on the initial hypothesis and found that, where present, absolute value markers did not significantly influence the final rating.

## Final ratings

186. Chart 3 below shows the profile of final ratings awarded at provider level. This broadly reflects the profile of initial hypotheses shown in Chart 1 above, with more providers awarded Bronze than any other rating, followed by Silver, then Gold. Therefore, although the distribution of Bronze, Silver and Gold ratings generated in this pilot were more heavily skewed towards Bronze than in previous provider-level exercises or in the previous year's pilot[34], this was driven by the sample of providers chosen to participate and their metrics starting points, rather than a methodology that generated more Bronze ratings in this pilot.

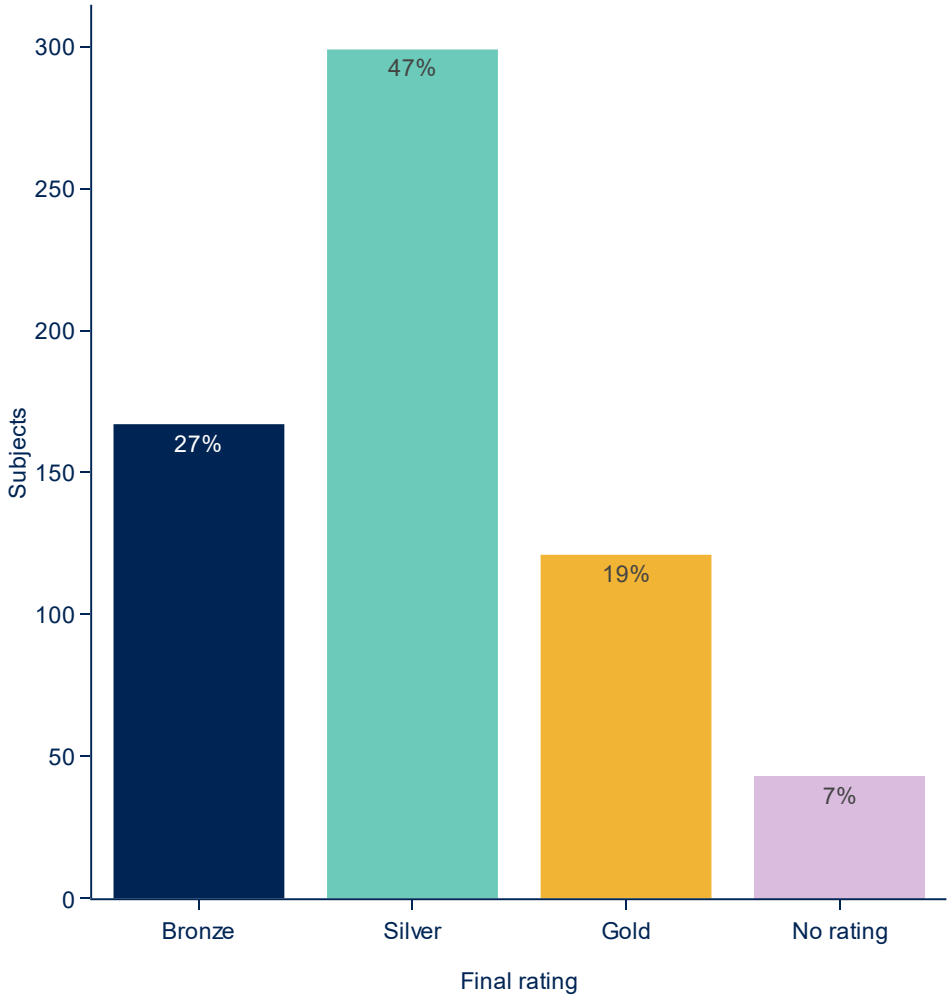
### Chart 3: Final provider-level ratings




187. The profile of subject ratings produced in the pilot is shown in Chart 4: Subject-level final ratings below. Around half of subjects were rated Silver, a quarter were rated Bronze and a fifth were rated Gold. This profile more closely matches the distribution of ratings generated in previous exercises. 43 subjects received no rating.

[34] Available at: https://www.gov.uk/government/publications/teaching-excellence-framework-lessons-learned; Findings from the first subject-level pilot 2017-18, available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-findings-from-the-first-subject-pilot-2017-18/.

### Chart 4: Subject-level final ratings




## Movement of ratings from initial hypotheses

188. Analysis of the movement of ratings through different steps of assessment shows that while the final rating often reflected the initial hypothesis, in many cases the final rating changed as a result of panel judgement. Chart 5 and Chart 6 below summarise the movement observed throughout provider-level and subject-level assessment processes, highlighting the overall rating that each provider or subject ultimately received. At both provider and subject levels:

    a. Movement away from the initial hypothesis was observed in both directions – to higher and lower ratings.

    b. No assessments moved from Bronze to Gold or vice versa. In subject-level assessment (see Chart 6 below), six subjects moved from Silver/Bronze to Gold, which was the largest movement observed – in four of these cases the inclusion of absolute values accounted for half a grade of the total movement. Movement was more often limited to an adjacent rating, although it was not purely restricted to cases which had initially started at ‘borderline’ ratings.

c. There was a greater tendency towards ratings being moved up than moved down. Table 12 below shows that this mirrored historical patterns observed in provider-level TEF.

189. Less movement was observed at subject level than at provider level, driven in part by a higher proportion of Silver initial hypotheses at subject level, which remained at Silver more often than was the case at provider level.

**Chart 5: Movement from initial hypothesis to final provider-level rating**

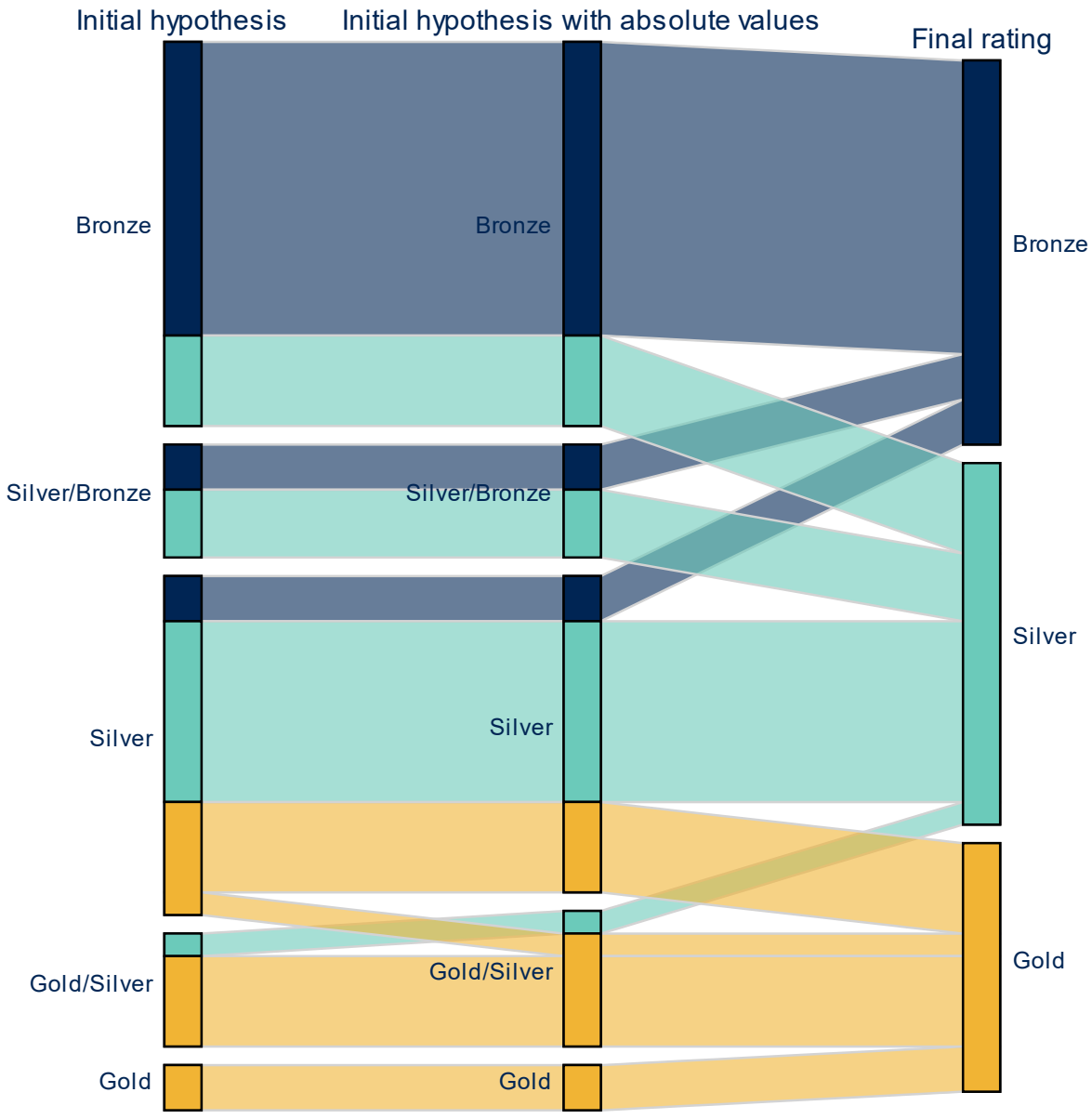
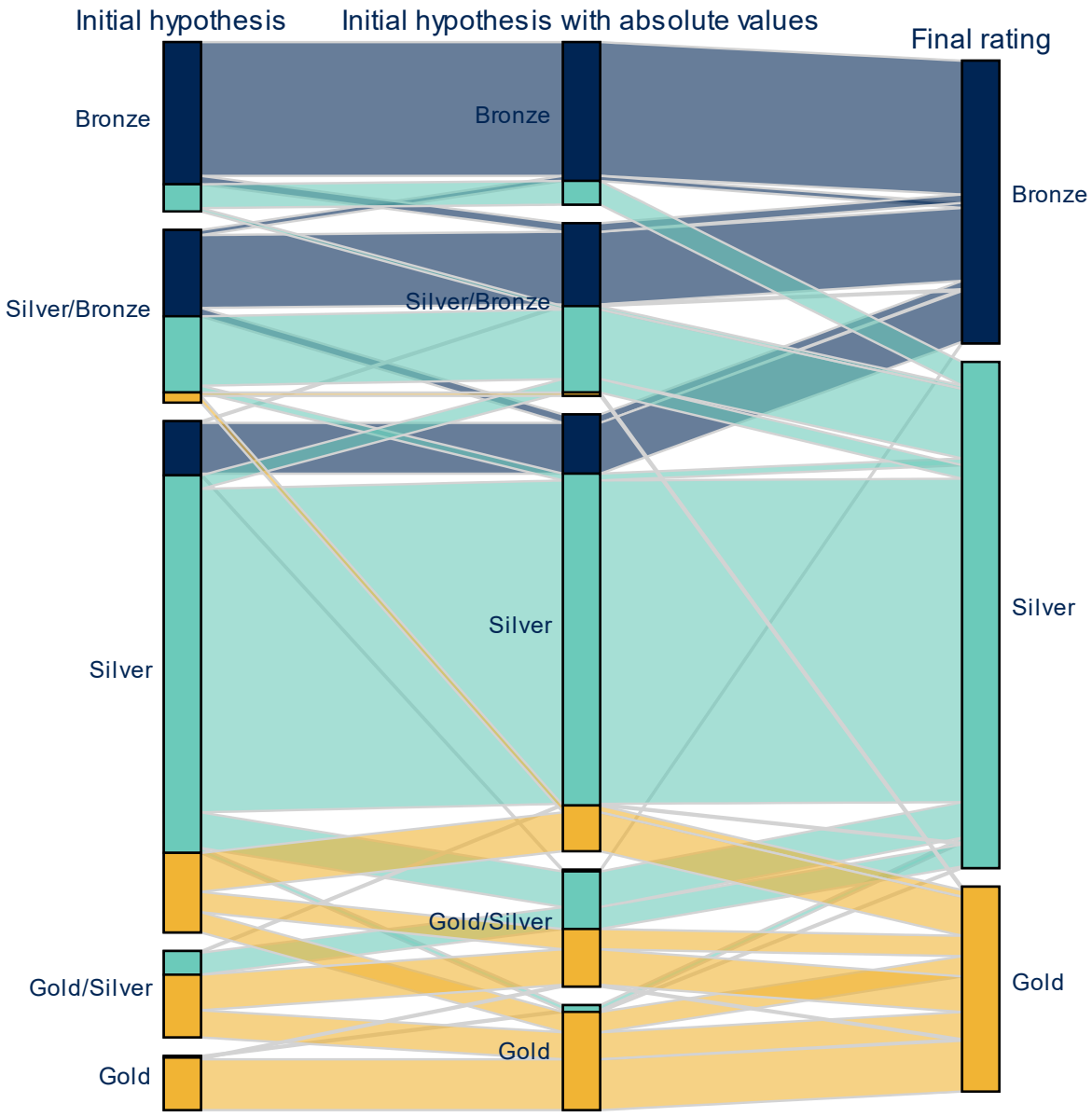


**Chart 6: Movement from initial hypothesis to final subject-level rating**



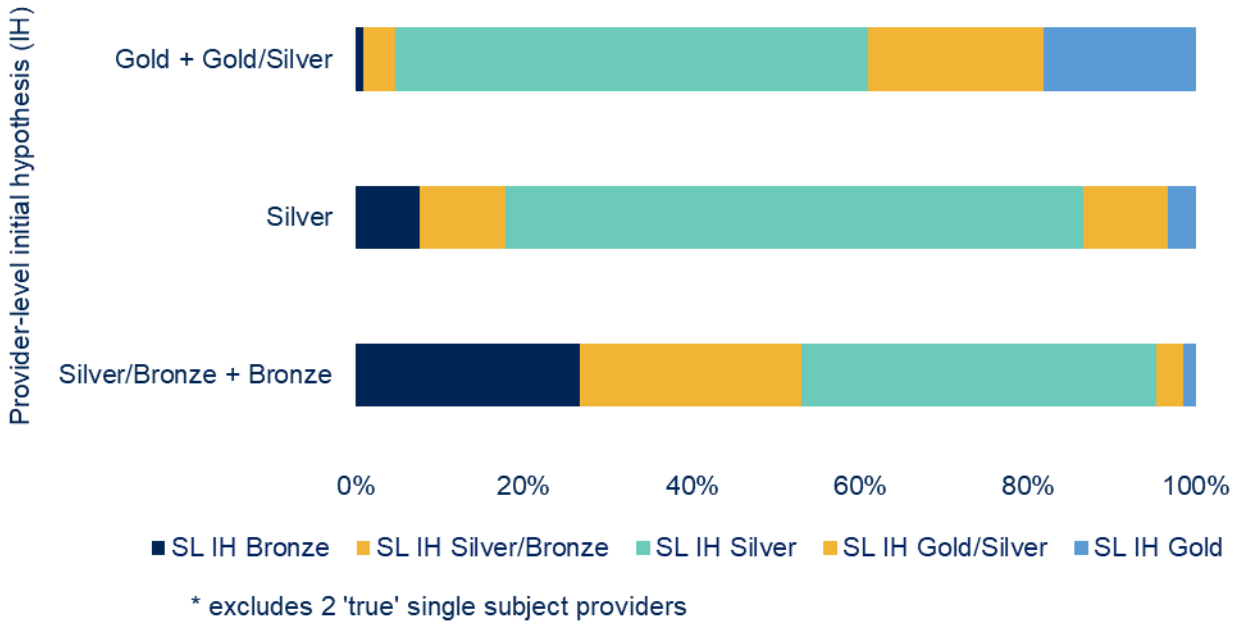
### Table 12: Movements between initial hypothesis and final rating in different TEF exercises[35]

| | 2018-19 pilot: provider level | 2018-19 pilot: subject level[36] | Provider-level TEF Years Two, Three and Four |
|---|---|---|---|
| **Moved down** | 5% | 6% | 9% |
| **No movement** | 75% | 83% | 71% |
| **Moved up** | 20% | 12% | 20% |

## Variation of subject performance within providers

190. Analysis of provider and subject-level step 1a initial hypotheses, across all pilot providers, reconfirmed the pattern seen in the previous pilot, that there is significant variation within providers when considering the metrics only. As Chart 7 below illustrates, almost every combination of provider and subject-level initial hypothesis was observed in the pilot. As might be expected, providers with lower initial hypotheses tended to have more subjects with lower initial hypotheses and similarly providers with higher initial hypotheses had more subjects with higher initial hypotheses.

### Chart 7: Differences between provider and subject-level metrics[37]



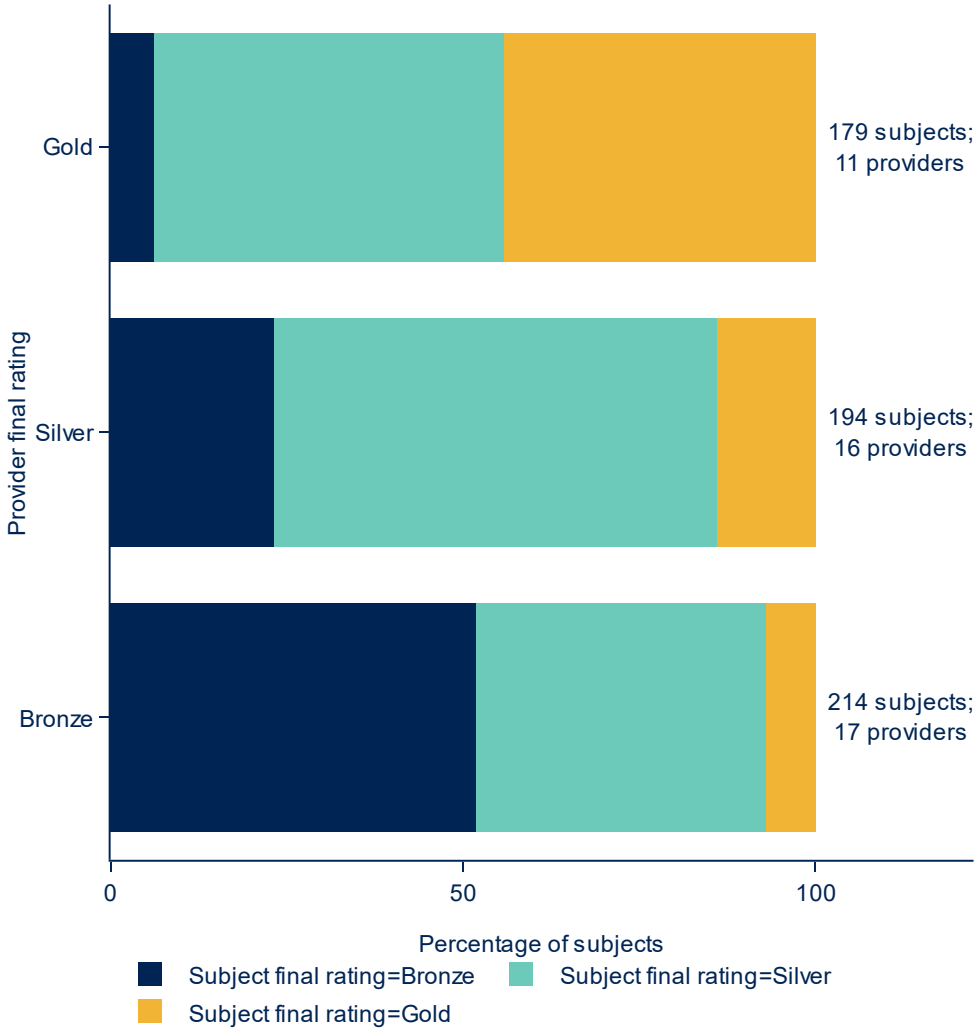

[35] For the purposes of this table, 'movement' in the 2018-19 pilot is defined as movement of a 'full grade' (e.g. Bronze to Silver), rather than movement required to resolve a borderline rating (Bronze/Silver to Silver).

[36] Percentages do not sum to 100 per cent because of rounding.

[37] Chart 7 groups together providers with Gold and Gold/Silver initial hypotheses, and groups together providers with Bronze and Silver/Bronze initial hypotheses to avoid disclosure of any individual provider's data.

191. These patterns continued through to the final outcomes. Excluding single-subject providers, only three providers (seven per cent) received the same rating at provider level as they received in every one of their subjects. In most cases there was at least some variation within a provider's profile of subject ratings and, for those providers for whom it was possible, almost two-thirds received at least one Gold, one Silver and one Bronze subject rating. The chart below shows the proportion of Gold, Silver and Bronze rated subjects belonging to providers receiving each provider-level rating.[38] [39] This shows significant variation in performance of subjects within providers, with both Gold and Bronze rated providers both having individual subjects that were rated at the opposite end of the scale.

**Chart 8: Differences between provider ratings and their subject ratings**




192. The exploration of aspect ratings also suggested that, in many cases, a single provider-level rating could mask underlying variation in performance across the three aspects. Where the panel tested the use of aspect ratings, there were also often variations between the overall rating awarded to the provider or subject and the more granular ratings produced for aspects of quality. Aspect ratings were only tested in a limited number of cases, but from this sample

[38] 'No ratings' are excluded.

[39] Note that this chart only counts subjects, and does not take into account the numbers of students in those subjects – a factor which was taken into account by the main panel in determining the provider-level rating.

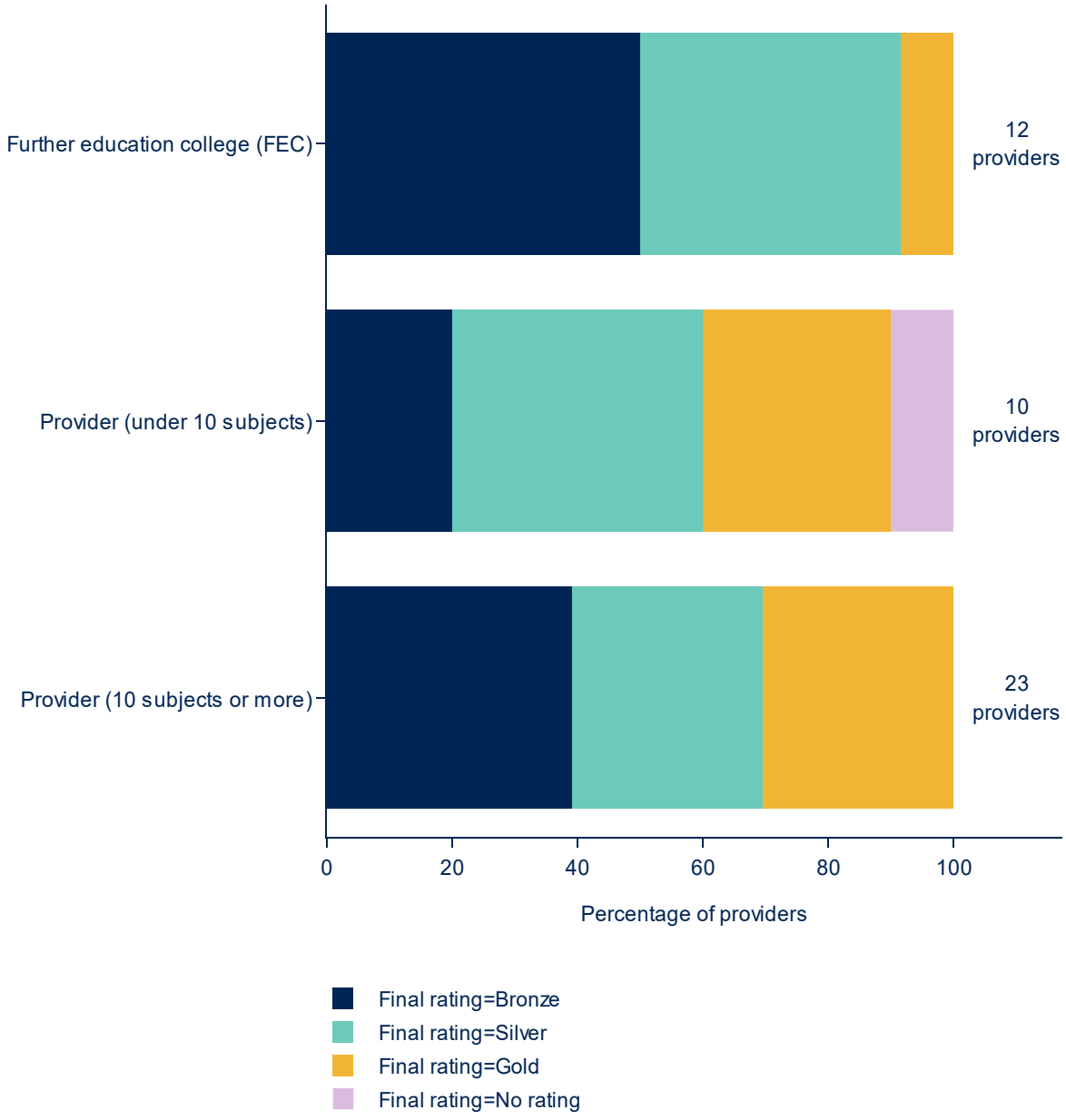
there were indications that the Student Outcomes aspect was most likely be rated differently to the overall rating.

## Ratings by provider characteristics

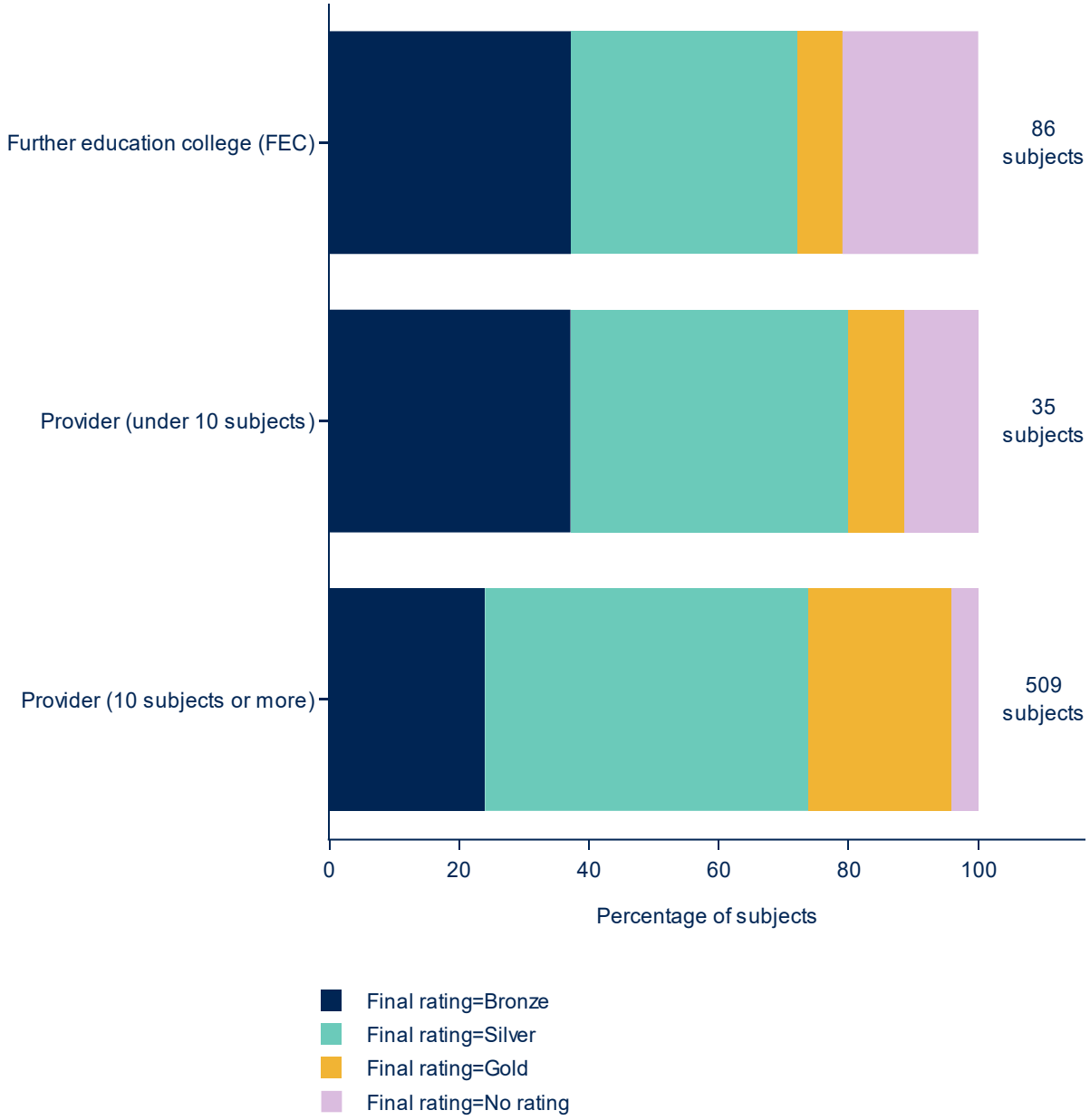
193. At both provider and subject level, all types and sizes of providers were awarded the full range of ratings. While provider type or size did not therefore act as a ceiling to the level of rating that could be achieved, the profile of these awards may suggest a differential impact for certain providers.

194. Chart 9 and Chart 10 below show the ratings awarded in the pilot by provider type. The categories used to define 'provider type' throughout this analysis are explained in paragraph 41 and Table 4.

### Chart 9: Final ratings by provider type



## Chart 10: Subject ratings by provider type

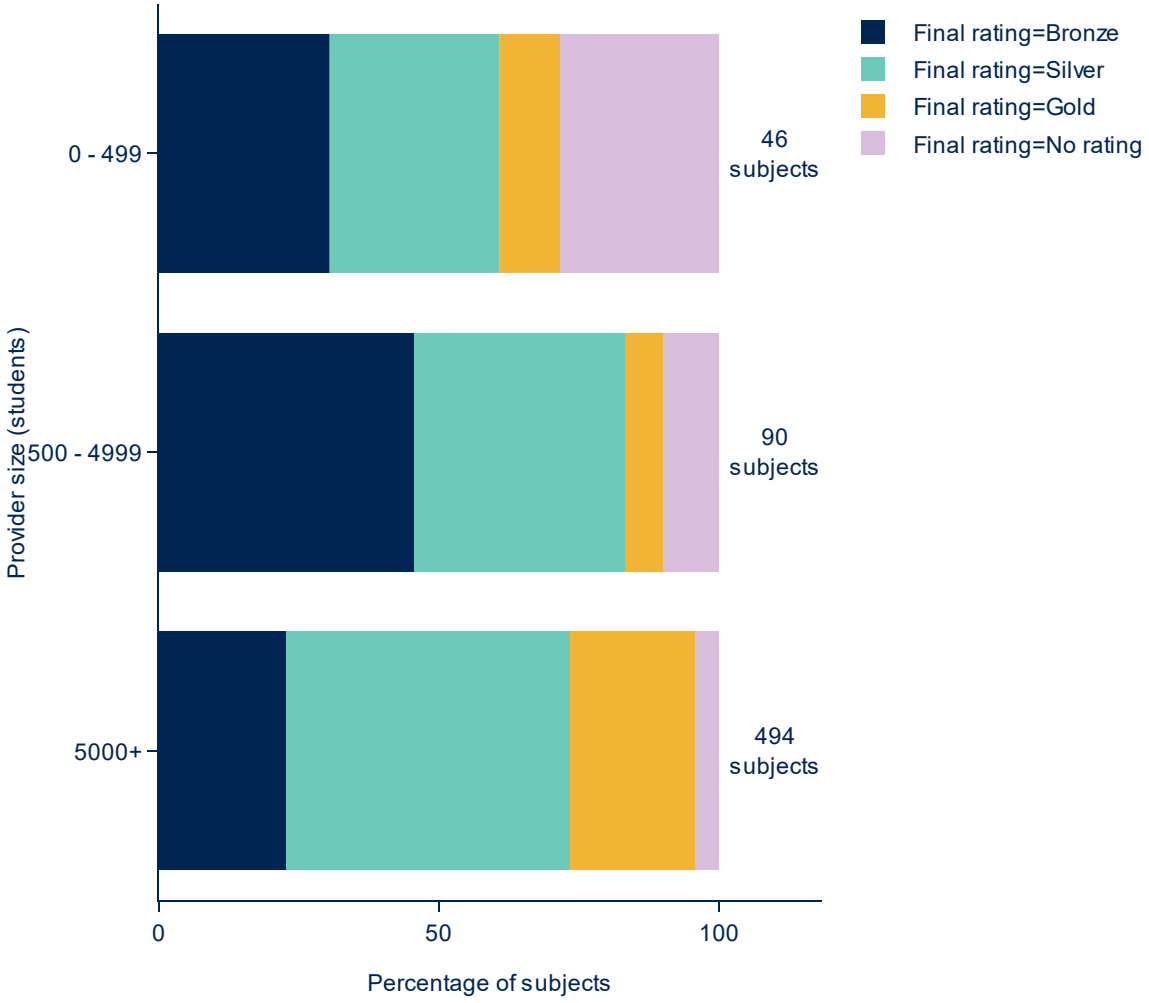



195. Chart 9 shows that in provider-level assessment, further education colleges were less likely to be rated Gold and more likely to be rated Bronze than other types of provider. At subject level (see Chart 10), both further education colleges and providers with a narrower range of subjects were less likely to receive Gold ratings and more likely to receive 'no rating'.

196. Regression analysis (see Annex G) also found that provider type was a significant predictor of subject performance, with subjects at FECs or providers with fewer than 10 subjects receiving lower ratings than those at providers with 10 or more subjects.

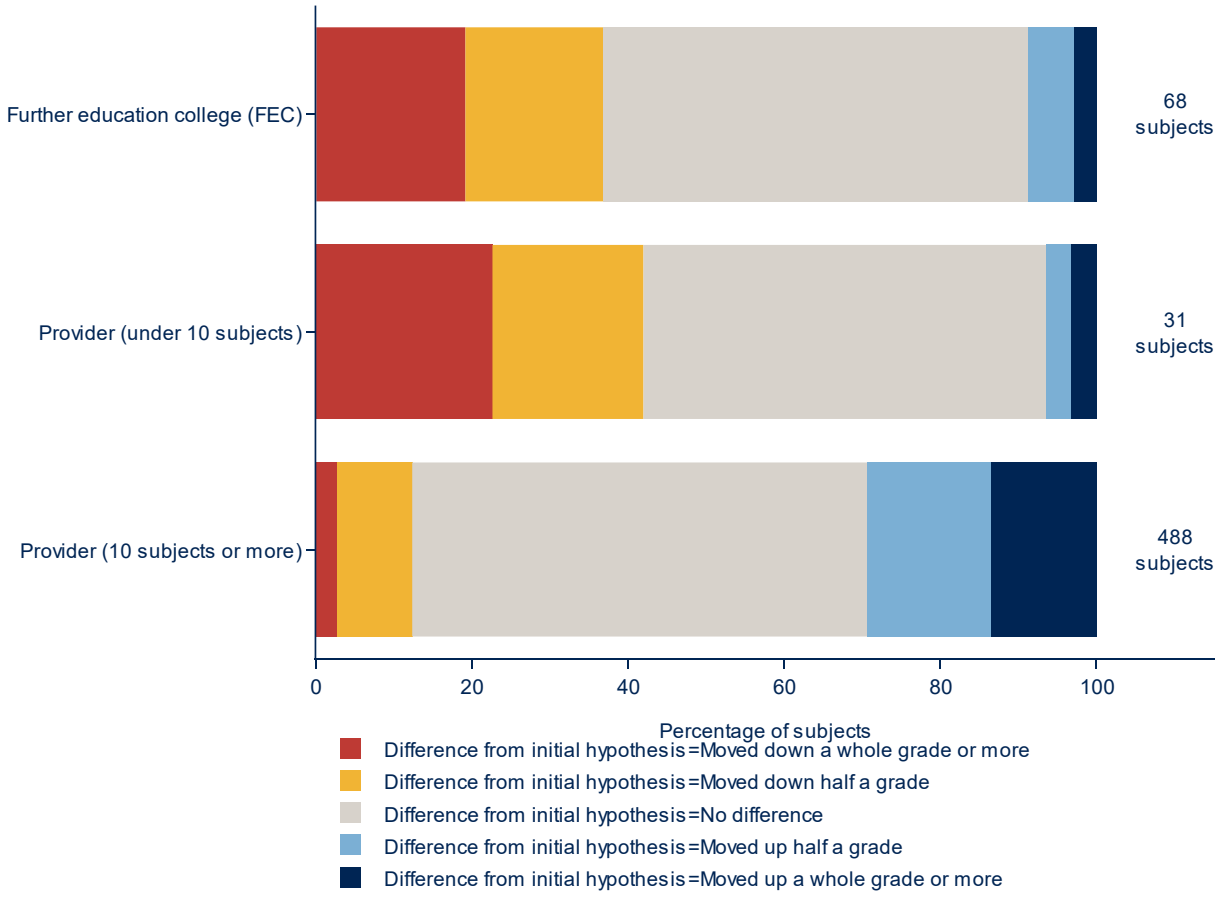
197. Similarly, smaller providers (based on student full person equivalent (FPE) were less likely to be awarded Gold and more likely to be receive no rating at both provider and subject level. Chart 11 below shows subject-level ratings by provider size. Provider FPE also appeared to have an effect on the subject-level rating when included in the regression model, but its impact was not significant when the provider-level award was included as a factor.

## Chart 11: Subject ratings by provider size




198. This differential impact also applied to the way providers were moved away from their initial hypothesis by the panels. While there were examples of all types of providers moving away from their metrics starting points in both directions, the distribution of movement was not even. Chart 12 below shows that in subject-level assessment, subjects at further education colleges or at providers with a narrow range of subjects were more often moved down from their initial hypothesis and less often moved up when compared to providers with a broad range of subjects.

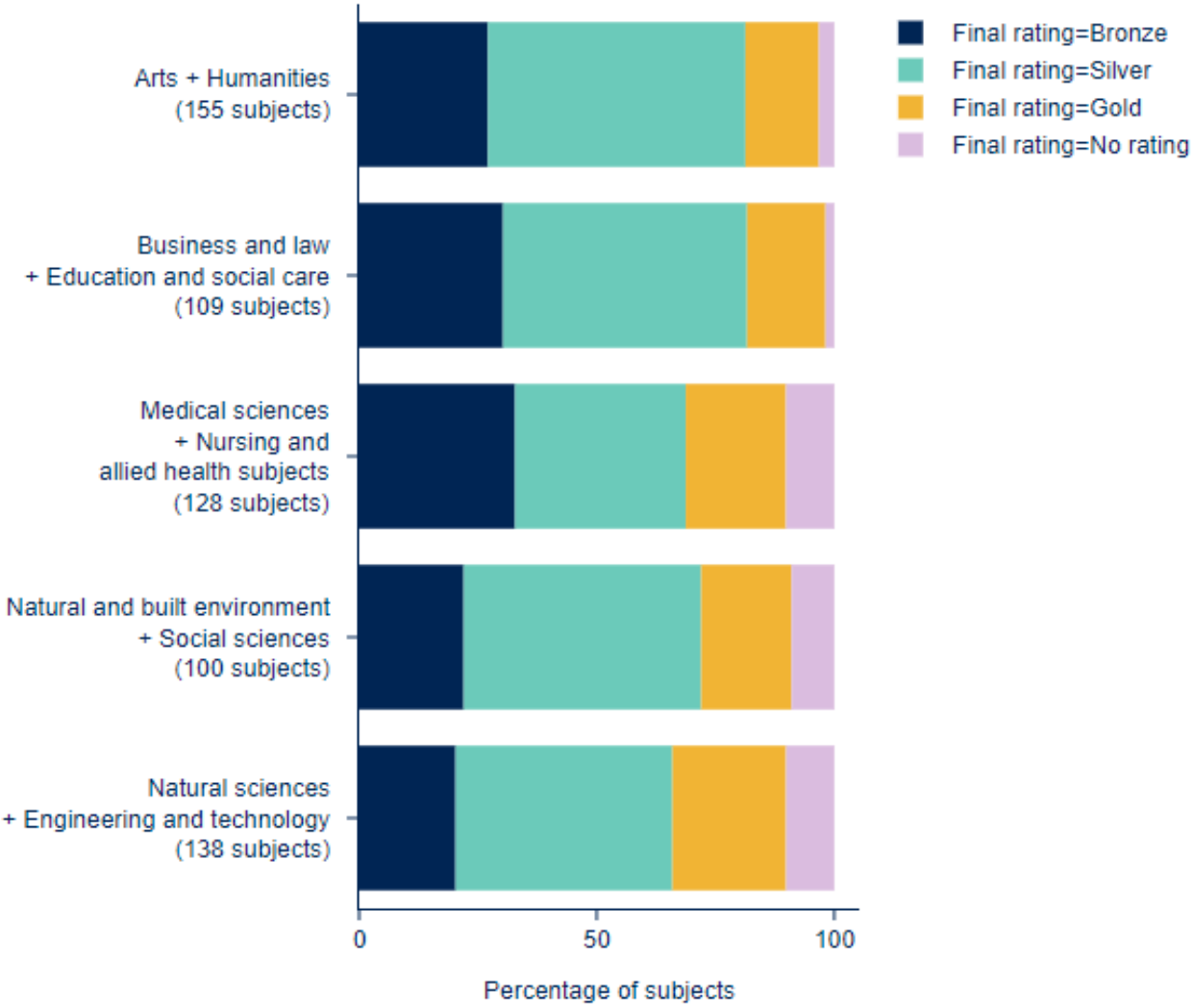
### Chart 12: Differences between subject ratings and their initial hypotheses by provider type



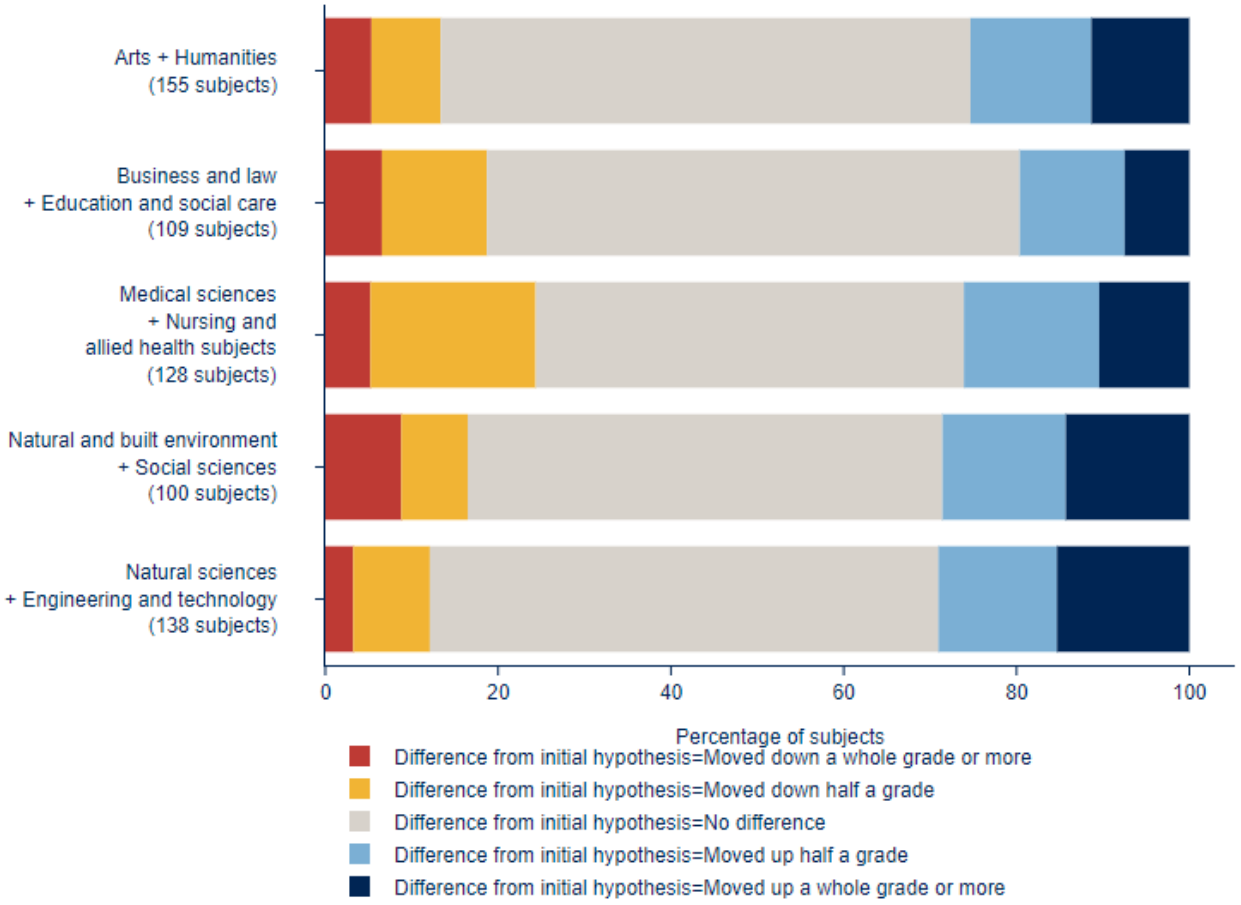

## Differences between subject panels

199. The overall profile of ratings arrived at by different subject panels broadly followed the overall pattern of subject ratings described above. However, there was some variation between subject panels in terms of the exact proportion of subjects awarded each rating, and the proportion of subjects each panel chose to move up or down (presented in Chart 13 and Chart 14 below). The OfS's regression models suggest in particular, taking into account all other factors, subjects assessed by the Medical Sciences and Nursing and Allied Health panel were likely to receive worse ratings than subjects assessed by other panels, and were the most likely to receive no rating. Furthermore, there were significant differences in the proportion of no ratings awarded by a number of the panels which were not explained by the metrics alone.

## Chart 13: Subject-level ratings by panel



### Chart 14: Differences between subject ratings and their initial hypotheses by panel



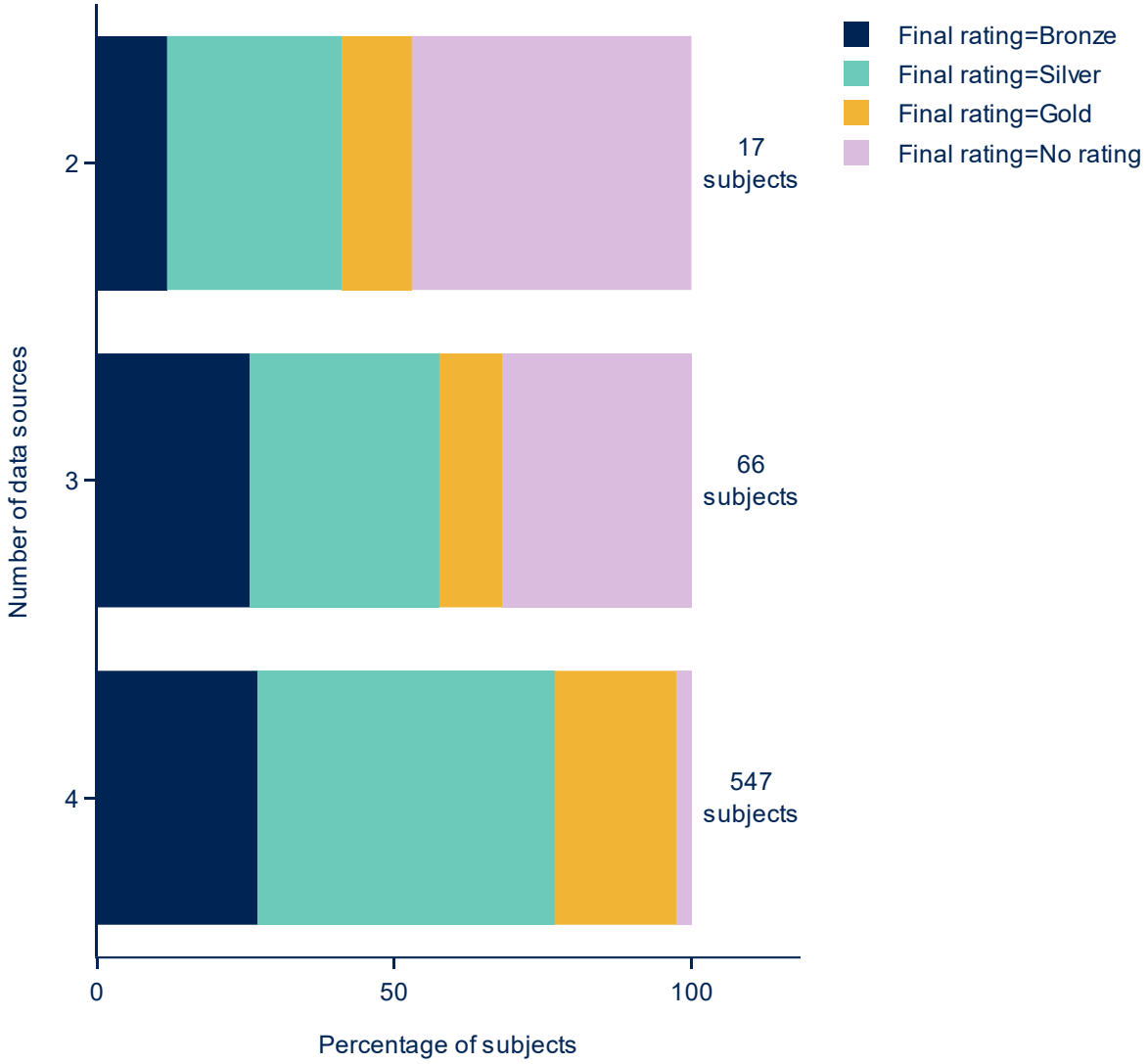

200. Panels set aside time during the course of assessment to reflect on these patterns and consider inter-panel consistency as they progressed through their caseloads. Although panels were confident that the assessment process had been consistently followed within each group, this highlights the significant challenge of moderating across multiple panels.

201. Each panel's caseload of subjects was divided into two batches, which were assessed at different meetings as a way of testing processes to facilitate delivery at scale. The OfS regression analysis found that batch of assessment was not a significant predictor of a subject's final rating, suggesting that panels managed to maintain internal consistency between meetings, even if consistency between panels was more difficult to achieve.

202. Some variation was also observed in the ratings awarded to individual subjects. However, small sample sizes, the impact of provider selection and panel structure make it difficult to draw any conclusions about the comparative treatment of subjects in the pilot or the performance of subjects across the sector.

## Data availability and limitations of subject-level data

### Chart 15: Subject ratings by number of available data sources




203. TEF metrics draw on four sources of data: the NSS and DLHE surveys, HESA or ILR data and the LEO dataset. Minimum data thresholds for this pilot meant that at least two of these sources had to be available in order for a provider or subject to be assessed. 83 subjects in the pilot had at least one missing data source. Chart 15 above compares the ratings produced for these subjects with those where all four sources of data were available.

204. A small number of subjects with missing data were able to achieve Gold ratings, despite the broader problems that missing data caused. Where panels found it possible to award a rating to subjects with missing data, they mirrored the general pattern of being rated Silver most often, with smaller proportions of subjects awarded Bronze and Gold. However, the likelihood of a subject not receiving a rating increases significantly when data sources are unavailable. Almost half of the subjects where two data sources were missing received no rating. Regression models which controlled for other factors corroborated the significance of this relationship.

205. In addition to the increased likelihood of no rating, analysis of missing data sources showed that:

    a. Data sources were more likely to be missing for certain types of provider. In the pilot, only eight per cent of subjects at providers with a broad range of subjects (typically large universities) had missing data sources, compared to 34 per cent of subjects at other types of providers.

    b. Across the sector, missing data sources would lead to a higher rate of Silver initial hypotheses, and fewer Gold.

    c. In the pilot, and across the sector as a whole, LEO data was the most likely data source to be absent, followed by NSS data, then DLHE data.

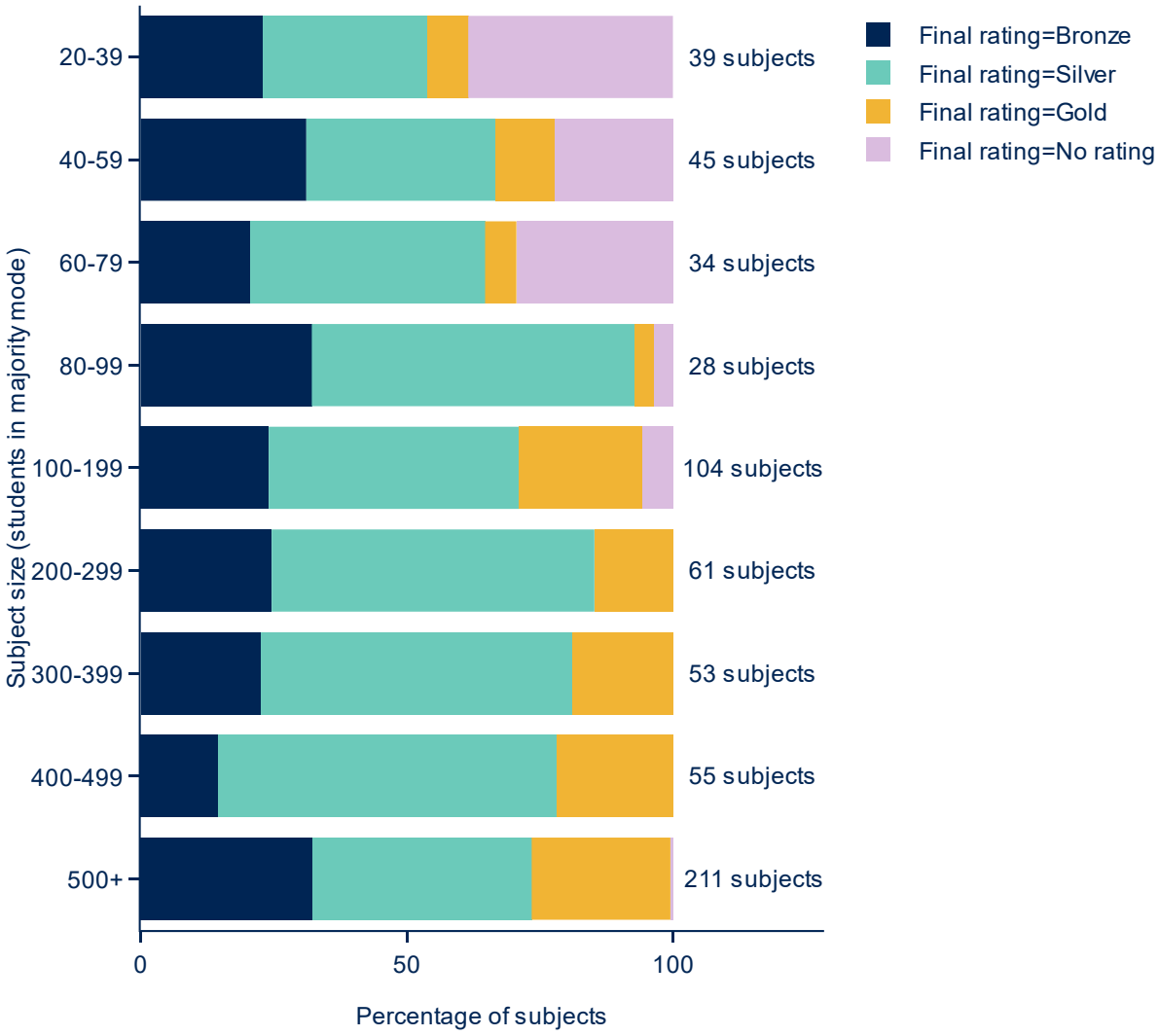
    d. Sector-wide analysis also shows that the pattern of more Silver and fewer Gold was stronger for subjects with missing NSS or DLHE data, compared to subjects with only LEO data missing. This may suggest the absence of LEO data had less impact on a provider's metrics than the absence of other metrics (even though its absence was more common).

206. While the impact of provider size has been considered above, subject size has been considered independently. Although there is likely to be some correlation between provider and subject size, subject size may be more closely linked to issues around data availability and statistical confidence in the data.

207. Chart 16 below shows that subjects of all sizes were able to achieve the full range of ratings.[40] However, smaller subjects were more likely to be awarded no ratings and were less likely to be awarded Gold. This pattern is particularly visible for subjects with fewer than 100 students.

[40] For this analysis, subject size is defined as the number of students (in FPE) in the provider's majority mode of study.

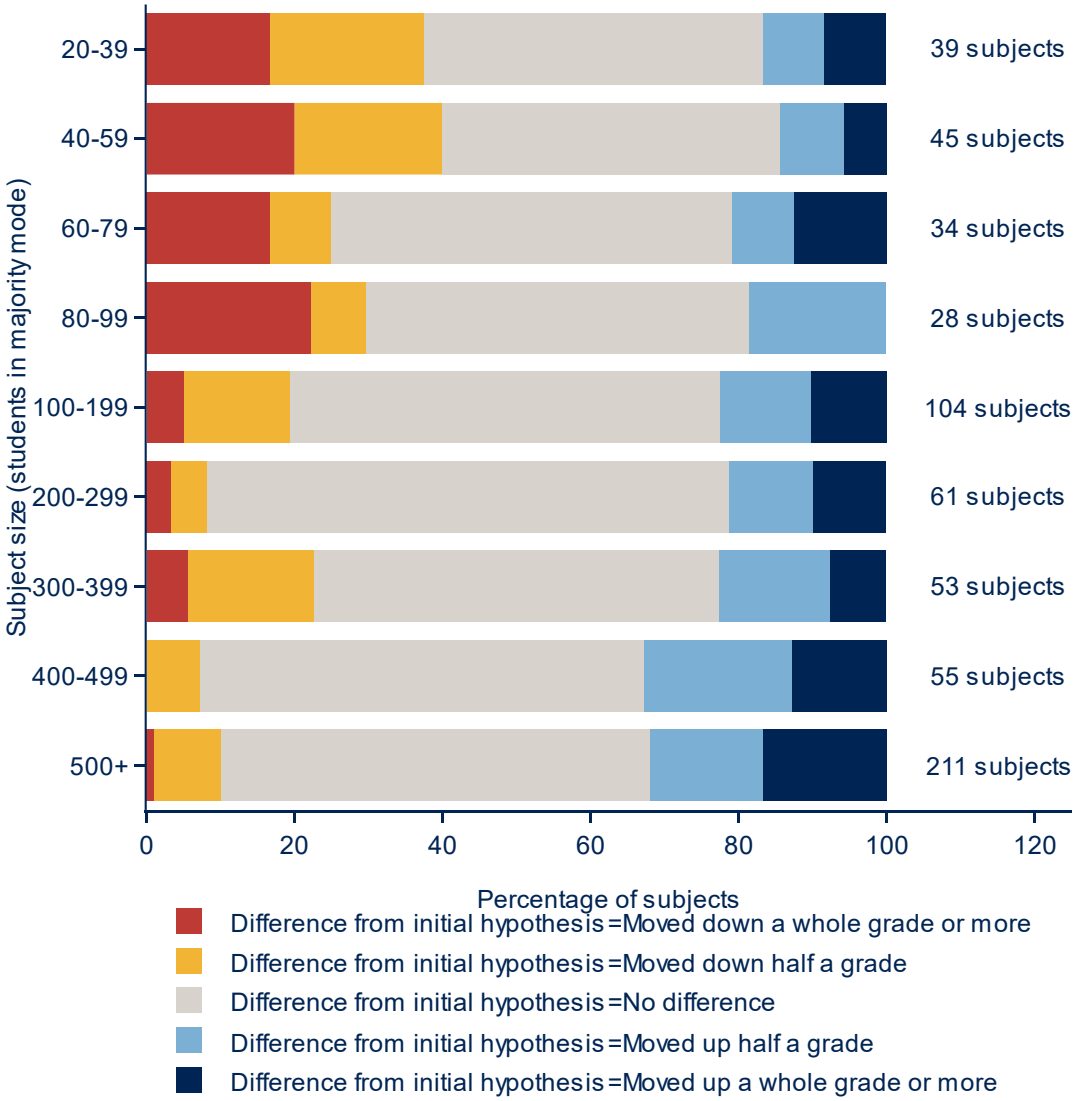
## Chart 16: Subject level ratings by subject size




208. Furthermore, analysis of panels' judgements by subject size shows that smaller subjects were much more likely to be moved down from their initial hypothesis to a lower final rating than larger subjects. The largest subjects (above 400 students) were also the most likely to be moved up.

## Chart 17: Differences between subject ratings and their initial hypotheses by subject size

209. Regression models confirmed that, controlling for other factors, smaller subjects were more likely to be awarded 'no rating' and larger subjects were less likely. However, the models did not find a significant relationship between subject size and the likelihood of a Bronze or Gold rating being awarded.

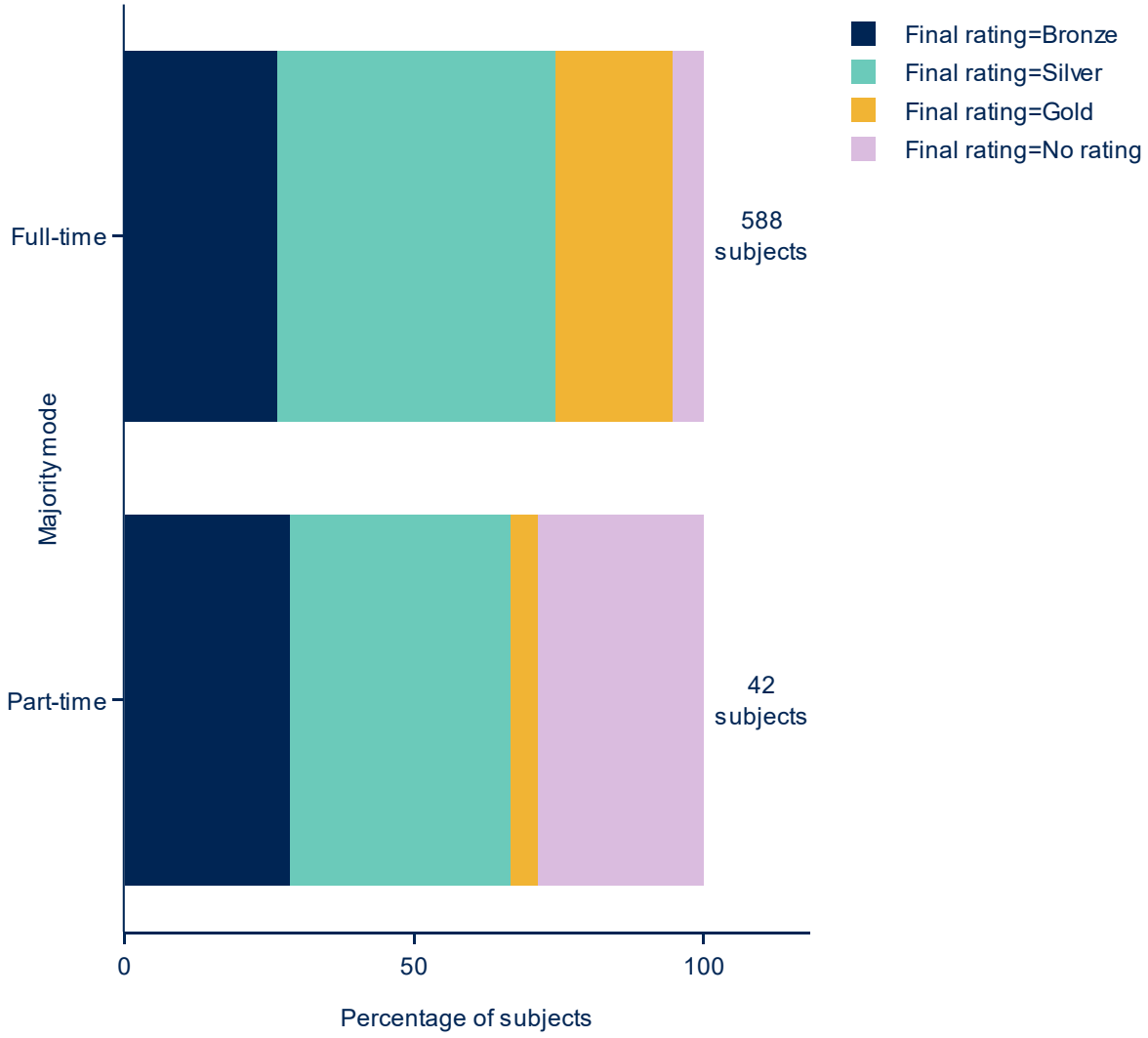


## Other subject characteristics

210. The vast majority of subjects assessed in the pilot were full-time in majority mode.[41] However, analysis of the ratings for the small number of subjects that were part-time in their majority mode suggests increased likelihood of part-time subjects receiving no rating and less likelihood of being rated Gold, as shown in Chart 18 below.

211. Regression models did not find majority mode of study to be a significant predictor when other factors such as provider type and initial hypothesis were controlled for.

### Chart 18: Subject-level ratings by majority mode




212. Analysis of subject ratings by the degree of interdisciplinarity within a subject was also undertaken.[42] While around a third of subjects assessed in the pilot were completely self-contained (that is, the metrics were based on data reported to that subject alone), the

[41] The majority mode of study is the mode of study in which the majority of a provider's students are engaged. This is determined by the proportions of students recorded in each mode of study in the contextual data of a provider's TEF metrics.

[42] Interdisciplinarity here is defined as the extent to which students are studying on courses which span multiple CAH2 subjects, and therefore data relating to them is split between multiple subject workbooks.

majority included students who were split across multiple subjects. Around 37 per cent of subjects assessed in the pilot shared data with three or more other TEF subjects.

213. While panels and providers commented on the difficulty of assessing highly interdisciplinary subjects (see paragraphs 139-144 above), analysis of the pilot outcomes did not show any clear pattern in the ratings awarded to subjects by level of interdisciplinarity. Other factors considered above, such as provider type and size, appear to have a far greater influence on the final rating than interdisciplinarity.

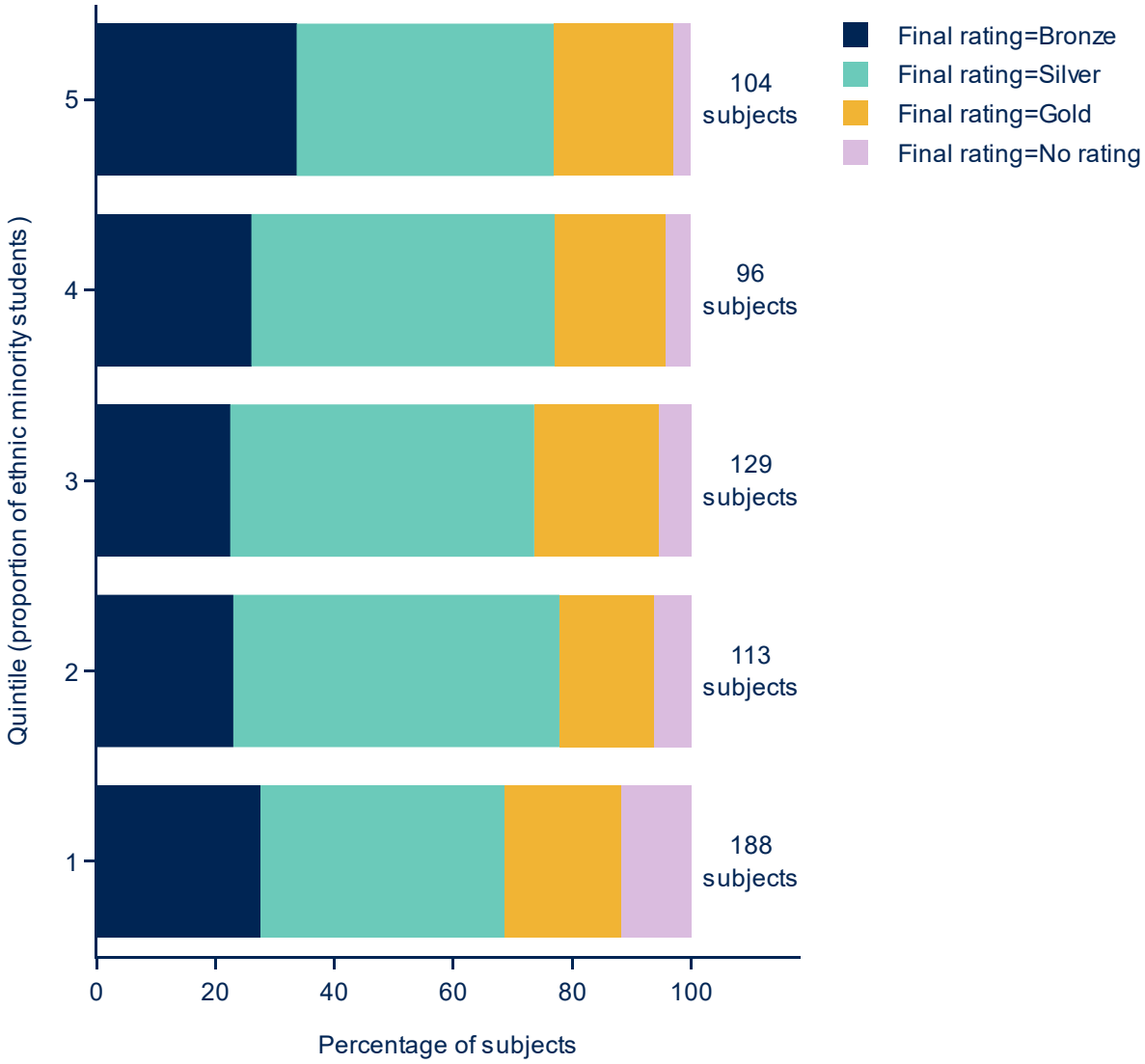
## Student characteristics

215. Analysis was also undertaken to investigate the patterns of ratings awarded to subjects according to the proportion of students with certain characteristics. This analysis looked at factors such as:

- age
- sex
- domicile
- level of study
- disadvantage
- entry qualifications
- disability.

To produce the charts below, subjects were divided into population-weighted quintiles based on the proportion of students in that subject with a given characteristic (quintile 1 having the lowest proportion of students with that characteristic, and quintile 5 the highest). For many of these factors, such as the proportion of black, Asian and minority ethnic (BAME) students shown in Chart 19 below, there was no observable pattern in the ratings awarded.

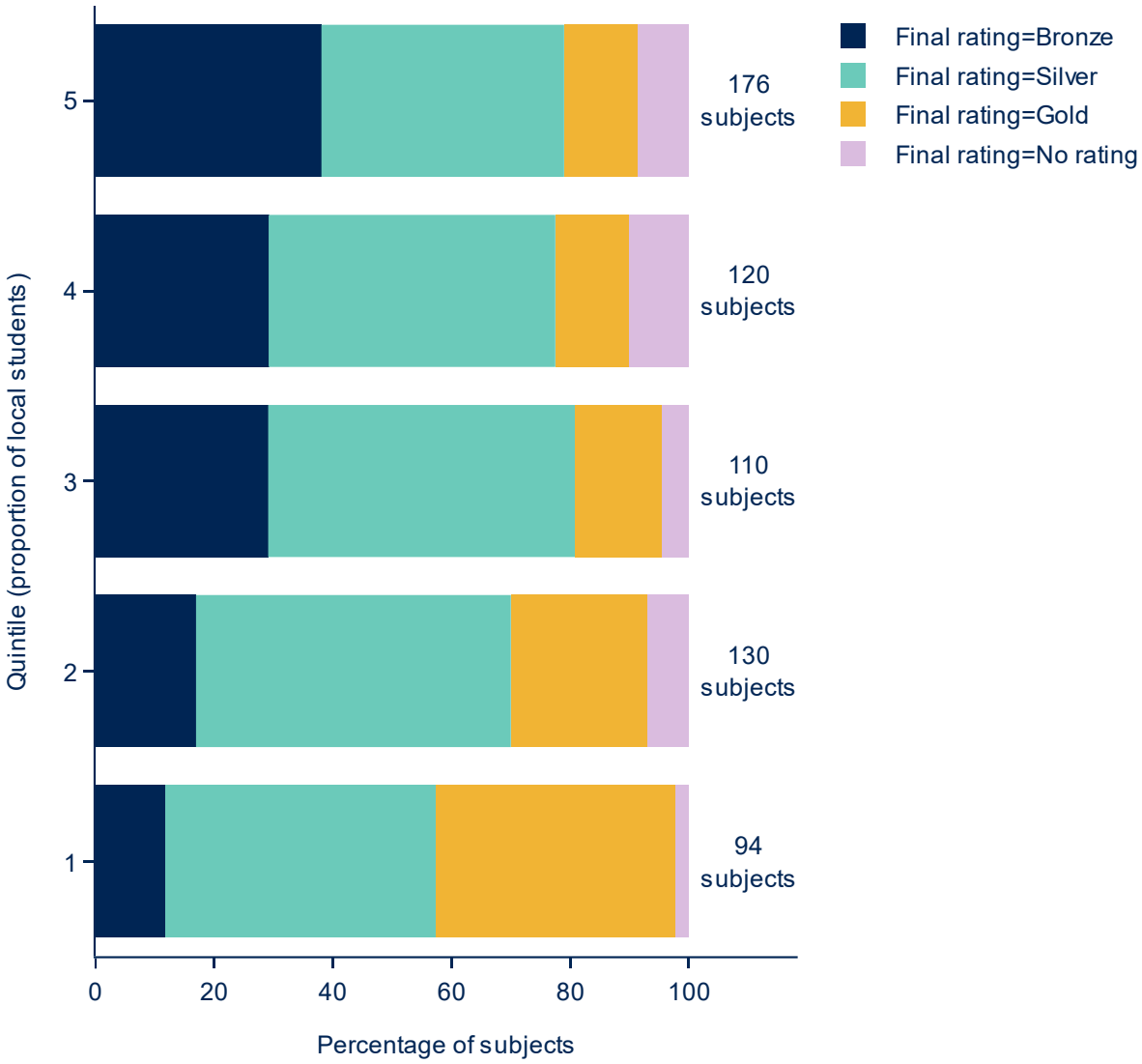
**Chart 19: Subject ratings by ethnicity (proportion of BAME students)**




216. For other factors there was a more discernible pattern, such as ratings awarded by the proportion of local students within a subject (shown in Chart 20 below).

217. Such student characteristics were tested as predictors in the regression models used to analyse the pilot ratings, but were found not to have a significant effect when other variables such as provider type and initial hypothesis were taken into account. Patterns such as the one observed in Chart 20 below may therefore be explained by provider type, and the student characteristics which may be more prevalent at those types of providers.

**Chart 20: Subject ratings by proportion of local students**




## Relative impact of metrics

218. The regression model also tested the impact of individual metrics on the final rating, beyond the formulaic impact they had on the initial hypothesis. The analysis found that flags on some, but not all, metrics had an additional significant effect on the final rating, after taking into account their role in the initial hypothesis calculation. This was notably true for four out of five of the NSS metrics, but only one out of the three employment metrics. This suggests that panels made greater use of the satisfaction metrics in formulating their holistic judgements, over and above their half-weighting in the initial hypothesis calculation. The ‘teaching on my course’ metric was found to have the greatest additional impact of all the metrics, suggesting it was highly valued by panel members.

## Key findings and overall conclusions

219. This section sets out our key findings and overall conclusions from across all the reports and sources of feedback.

- **The comprehensive model piloted in 2018-19 was an improvement on the previous year's pilot models.** The revised model was comprehensive in that it ensured that all subject ratings were based on a full assessment of that subject. This addressed design issues with previous models, where some ratings were made with limited evidence or no direct assessment. However, the precise relationship between subject and provider-level ratings would need further consideration.
- **The pilot revealed significant variation in performance within providers, confirming the importance of taking account of this variation in the TEF outcomes.** It was common for individual providers to receive the full spectrum of ratings (Gold, Silver and Bronze) across their subjects. Scrutiny at subject level identified areas of excellence and areas for improvement within providers, in a way that assessing only provider-level evidence does not.
- **There was a clear expectation that subject-level TEF would help drive improvements in higher education.** Participating providers, students and panels reported that drilling down into subject-level performance would have positive benefits and incentivise excellence.
- **The TEF needs to be better aligned with other OfS functions.** Participants reflected that the TEF would have more internal impact if processes were more closely linked to other regulatory functions (in particular access and participation plans and ongoing monitoring against the other quality and standards conditions of registration). This would also enable providers to respond more efficiently. Better alignment would also help clarify how to interpret information that is relevant across these different functions .

220. In the first pilot, we reported how broad support for a more comprehensive model quickly gained traction with all groups of participants. Student panel members and students from providers 'felt strongly that a model which fully assesses each subject would produce a more meaningful and accurate set of ratings for applicants and students' whilst providers and panel members noted the approach would be fairer, more transparent, more representative, and overall more positive for institutional enhancement and engagement. Commentary of this can be found in the 'Findings from the first subject pilot'.[43] These views arose because it was clear that fundamental problems with subject-level assessment in the first pilot had been introduced by design: in both model A and model B, usable evidence was effectively withheld from the process.

[43] Findings from the first subject pilot, page 22. Available at: https://www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-findings-from-the-first-subject-pilot-2017-18/.

221. In this pilot, where there was evidence available it was always assessed. Every subject had its own set of metrics, its own separate submission and was reviewed fully by the relevant panel. In practice, the change to this comprehensive model was found to be positive and gained little attention, though it is important to note that complexity and burden remained common concerns, and these are discussed in more detail below. Panels were more confident in applying the framework (see paragraph 111) and where there were barriers to making rating judgements at subject level they typically arose from the structural and statistical data issues.

222. Additionally, a wider range of framework refinements were successfully tested. In particular, having distinct processes and criteria at both subject and provider level increased the panels' confidence and ability to make judgements at each level (see paragraphs 111-114). Assessment of provider and subject-level information was not completely independent and further work on the assessment process would help to ensure that outcomes at both levels for a provider do not diverge in a way that is not meaningful or explainable for prospective students (see paragraphs 149-153).

> "In principle this model – in which all of a provider's subjects are assessed individually – worked well. Subject panels felt well-equipped to make assessments of subjects, and the subject panel structure and process made it possible to assess subjects in a logical framework. A mixture of academics, students and employer and professional, statutory and regulatory body (PSRB) representatives brought fresh insight and deep expertise which helped provide the nuance that may otherwise have been lost in an exercise operating on this scale."
>
> Main panel chair's report

223. Analysis (see paragraphs 190-192) illustrates that providers in the pilot had subjects where performance was different to the provider as a whole, and also that they could have a great deal of variation between the subjects they offered. Indeed, an implication of the analysis of Type II error rates (see Annex H) is that there is more variation than is accounted for in the current flagging method. The pilot demonstrated how providers were beginning to focus in on these areas for improvement. The IFF report highlighted that providers planned to enact a range of actions as a result of participation in subject-level TEF, such as using the results as a basis for future improvements, strategies and action plans, and embedding subject-level TEF into existing quality assurance and teaching enhancement processes.

> "We altered our programme enhancement processes to map onto the TEF data."
>
> TEF main contact (University)
> TEF subject-level pilot evaluation – Provider perspectives report

224. While often these activities were not yet having an impact in the timeframes of the metrics being assessed, there was a sense of optimism from panels during assessment that many of the activities put in place would have future positive impacts. There was also a sense that, as the process matured, the content of submissions would represent a valuable sector resource that would reveal a wealth of excellent practice.

225. Various participants saw added benefit from moving from provider to subject-level assessment, though their appetite depended on what they perceived the key purpose to be. Student contributors in providers were particularly keen as they foresaw improvements to student information. The majority of student contributors (89 per cent) considered subject-level TEF to be more useful than provider-level TEF. On the other hand, providers weighed their views against the perceived burden. They also felt that a future iteration of TEF should be more closely linked to other regulatory and enhancement mechanisms, in particular access and participation plans and ongoing monitoring against the conditions of registration.

"If you want to have the best provision possible, you have to invest in it and I think that this is a really important amount of time, and I don't think it's overly onerous."

TEF main contact (University)
TEF subject-level pilot evaluation – Provider perspectives report

"I felt that the process itself was streamlined, but the timescales and time period were challenging. It was very difficult to get students involved at a time of year that was close to Christmas and up against the January examination period. A longer gap between the release of the metrics workbooks and the narrative due date would have been beneficial."

Academic contributor (University)
TEF subject-level pilot evaluation – Provider perspectives report

- **Participants engaged constructively.** As they worked through the various processes, participants helped us to identify and explore a number of challenges involved in subject-level assessment. They engaged constructively in seeking to work through them, and whilst some issues remained unresolved, a great number of their ideas and suggestions helped to solve problems and refine processes.
- **Improvements were made to student engagement in the process and there was desire to extend this further in future**. The increased focus on student voice and student partnership in the assessment was welcomed. We trialled ways of increasing student involvement in TEF submissions and some progress was made particularly at provider level. However, there were practical barriers such as time and resource constraints, and involvement at subject level was patchy. Across panels and student representatives there was a desire to address these barriers, and further extend the opportunities for students to submit evidence more directly into the process.

226. As with the first pilot, we deliberately set aside time to refine some of the less well-developed processes in real time. Good examples of this include:

    a. developing the format and role of the student declaration through close engagement both with student representatives at providers (see paragraph 48) and student panel members (see paragraphs 115-118)

b. testing draft statement of finding processes at the calibration exercise, which allowed us to incorporate feedback into final guidance (see paragraph 86).

227. Provider TEF contacts were generous in sharing their knowledge and experience of working on TEF submissions at the various preparation events, with many colleagues contributing content to workshops which were always well-attended (see paragraph 48). There was positive engagement through all of the IFF survey work. Provider contacts were constructive on the occasions we asked them to try things that we knew were problematic (such as writing submissions where subjects were in scope but there was still very limited data) to make sure we were thoroughly testing where we had set thresholds (see paragraph 53).

228. All panel members, as the chair of the main panel reflected, had to 'hit the ground running' and worked with enthusiasm and professionalism, noting their huge contribution to the process:

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

229. All participants frequently went above and beyond in spending time with or corresponding with OfS colleagues to help trouble-shoot issues. Students in particular worked with the student deputy chair to arrange their own extra sessions to gather feedback and test ideas, and in panel meetings challenged panel conventions and assumptions (see, for example, paragraph 146).

230. We tested various mechanisms to increase the student voice, including:

a. separation of the TEF criterion 'TQ1: Student engagement' into two distinct criteria

b. introduction of new NSS metrics on student voice

c. introduction of TEF lead student representatives

d. testing of a student declaration

e. introduction of a student deputy chair role on the main panel.

231. Across all the evidence gathered by the pilot, we see how these developments were welcomed (see paragraphs 59, 115-120). But there was also a strong message that there should be greater ambition, in terms of creating as wide a range of opportunities for student input as possible and for a wider range of student-based evidence to be incorporated more fully into assessment. The IFF report (see Box 1) found that, in practice, there are numerous barriers to student engagement, especially at subject level. It is essential that these barriers are addressed and that the student voice is a clear and central tenet of the future framework.

"Overall, the main panel concluded that, while the concept of an independent declaration completed by students and submitted alongside the provider submission was useful to the assessment process, the declaration used was limited in that it could only establish a student's involvement in the submission process, and did not contain any narrative on how involved students were with their teaching and learning in a much wider context."

Student findings report

- **Pilot providers and panels encountered significant limitations in the data at subject level.** These included:
  − metrics that were missing or unreportable
  − data that was based on small cohorts and with limited statistical reliability
  − data that no longer related to the courses being offered in a subject.

  We tested means of mitigating these limitations but found that they did not resolve the issues. The panels needed a disproportionate amount of time deliberating over cases with limited evidence; they made judgements in each case on whether there was sufficient evidence to determine a rating, but this process was inconsistent and would not be scalable.

- **Further analysis shows that many subjects across the sector do not have large enough student cohorts for the current metrics flagging method to robustly inform assessments.** Analysis shows that under the current method for generating flags and an initial hypothesis, a subject would need to cover several hundred students in order for the metrics flags to robustly inform the assessments. Many subjects across the sector have smaller cohorts than this. While panels had a general awareness that metrics based on smaller cohorts had limitations, they conducted their assessments without knowledge of how large the cohorts needed to be to consistently generate flags, as this analysis was only carried out after the assessments had concluded. This could be addressed, though, through further work to improve how metrics could be generated and used robustly at subject level.

232. To help address issues around small cohorts and missing data, this pilot introduced minimum data thresholds, and tested whether panels could make more use of available metrics data in such circumstances, or whether submissions might be able to mitigate data limitations and enable robust assessments.

233. The high proportion of 'no rating's awarded to smaller subjects or subjects with missing data suggest the additional measures were not sufficient to remove uncertainty around assessments for such subjects. In particular, analysis of ratings by subject size supports the finding from the first pilot that a higher threshold for the minimum number of students would be required in order to consistently generate subject ratings with confidence (see paragraphs 203-209).

234. The analysis of pilot outcomes described above suggested that panels struggled to consistently determine ratings for subjects with fewer than 100 students. However, as

explained in paragraphs 173 to 183, additional analysis of the statistical power of the data for different sized cohorts suggests there were significant limitations in the ability of the flagging methodology to accurately identify actual differences in provider performance for cohorts which were much larger than 100. Panel members would have had confidence that the unflagged data was showing expected performance, but subsequent analysis shows there was actually a strong likelihood that good and poor performance was being missed due to the limited statistical certainty that the correct flags had been generated. To resolve this disparity, panel members would need to be given more information about the statistical uncertainty which underlies the presence or absence of a flag.

235. Methodological adjustments to try and address data limitations, such as the use of the 90 per cent confidence intervals and imputation of provider-level data for missing subject metrics, were either so rarely commented on by providers in their submissions, or treated with such little confidence by panel members, that they had little impact on assessment other than adding to the complexity of the process and density of information to be considered.

236. Panels did find that evidence in the submission could be used to effectively mitigate missing data, but noted that this was inconsistently done and expressed concerns about relying on this approach at scale. Mitigating actions would need to be put in place to ensure that, where assessments rely on additional provider-submitted evidence alone, it does not advantage those larger providers that are able to dedicate significant resource to submission writing, and disadvantage smaller providers that are not.

> "The panels felt that some providers were able to offer good mitigation by including a strong narrative and extra internal data within the submission, but this was not consistent from case to case. The majority of cases with a paucity of data failed to convincingly fill the gap. Panels found that when evidence was scarce, pieces of information that might otherwise have not had a large influence on the outcome gained prominence, but this was not always a fair approach."
>
> Main panel chair's report

- **Improvements to the subject categories following the previous year's pilot were welcomed, and some further refinement would be needed in this area.** Although improvements were made to the subject categories, some mismatches with internal provider structures remained. Where provision did not map to the subject categories, providers experienced difficulties in interpreting their metrics and writing submissions, and the ratings they received were of limited value for enhancement. There were also concerns that fixed subject categories would constrain innovation. While some providers suggested they should have flexibility in categorising their subjects, no solutions have been proposed at this stage that would be scalable. Further work would be needed to assess the scale of the mismatches and to identify further improvements.

- **Concerns remained about the treatment of interdisciplinary provision.** Only one-third of subjects assessed in the pilot were completely self-contained (i.e. with all students mapping to a single subject category). The majority included students studying in multiple subject categories, and around 15 per cent of subjects included students also studying in five or more other subject categories. We trialled the use of additional data and expertise within panels which helped to take account of interdisciplinary provision. While this was found to be helpful, challenges remained.

237. The revisions to the CAH2 subject classifications tested in this pilot represent the best possible compromise of different sector views on how to define subjects for TEF, but there are still areas of significant disagreement, both conceptually and logistically, which cannot be resolved in a single uniform subject classification (see paragraphs 59 and 98).

238. CAH2 also represents a compromise in terms of the level of granularity at which to assess subjects, meaning that for student information purposes, subjects assessed in TEF can sometimes represent a meaningless aggregation that bears little relationship to the programmes that students wish to apply to.[44]

239. Two-thirds of subjects assessed in the pilot drew data from courses which were coded to multiple TEF subjects, meaning that student data for those courses had been split between multiple workbooks (see paragraph 212). 71 percent of students were attributed to these subjects in the pilot. Often, panels and providers found dealing with such subjects extremely difficult and burdensome, in addition to the challenges this posed for student information.

240. Further work would need to be undertaken to fully quantify the extent to which subject mismatches occur across the sector more broadly.

"[A] common theme emerging in the second pilot was subject categorisation issues. Although some panels found the new CAH2[45] subject structure (revised for the pilot) helpful, most panels had at least one CAH2 subject where the level of aggregation was challenging […]"

Main panel chair's report

"We wrote three subject-level TEF submissions based on three individual programmes and then we wrote one subject-level TEF submission based on nearly twenty programmes. It was just… to write five pages around twenty programmes is really hard… if you've got areas that have got a lot of submissions, maybe they could be allowed a few more words."

TEF main contact (University)

[44] TEF and informing student choice: subject-level classifications, and teaching quality and student outcome factors, page 9. Available from: www.gov.uk/government/publications/teaching-excellence-framework-and-informing-student-choice.

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

"One in eight (13 per cent) of student representatives felt that subject-level TEF was less useful, due to the subject groupings being too wide: 'Because the subjects are too widely based, they are not specific.'"

Student representative (Specialist University)
TEF subject-level pilot evaluation – Provider perspectives report

- **The exercise was complex, and we identified some data that could be omitted in future.** We piloted additional metrics and contextual data, most of which were welcomed. However, the overall package of information became unwieldy and the metrics involve inherent complexity, requiring users of the data to be trained and to develop specialist TEF expertise. This made it challenging for providers to involve people across their departments, and challenging for smaller providers with limited capacity to develop this expertise. Tailored guidance for different roles within providers was suggested. Panel members faced a steep learning curve. They found that some of the data was of little use and could be omitted. In particular, the employment maps and contextual data could be omitted if regional employment factors were to be taken into account within the metrics.
- **The metrics relating to employment outcomes and differential attainment need further development.** There was a broad view that, if used, Longitudinal Education Outcomes (LEO) metrics should in future take account of regional factors. The pilot tested how experimental 'supplementary' data indicating gaps in degree attainment between groups of students, could be used in provider-level assessment. The main panel agreed that differential attainment could be considered within the TEF, as a key issue to be addressed in the sector. However, the panel found that to make use of the data in assessment, it would need to be developed further into a metric. Also, further clarity would be needed on how it relates to our regulation of access and participation.
- **The exercise demonstrated that more time would be needed throughout the process.** The complexity of the exercise was exacerbated by the tight timescales. This hampered the ability of providers to brief staff, analyse data, write and coordinate submissions across departments, and involve students. The panels achieved broad consistency in their decisions, but were unable on the pilot timetable to fully moderate their judgements. More time would be needed in a full exercise for all participants: for providers to write submissions, for students to engage in the process, for panels to be trained and conduct their assessments, for calibration, moderation and the consistent production of statements of findings, and for the OfS to develop systems, process data and deliver the exercise at scale.

241. From the outset it was our intention to use the pilot to test as wide a range of refinements as possible, with a view to streamlining the process in future. However, the pilot has demonstrated that there are more fundamental complexities and numerous tensions arising from the processes and data.

242. In exploring barriers to engagement, the IFF surveys identified a number of issues that illustrate how these complexities created difficulty in writing submissions: time, ability to interpret the metrics, having sufficient resources to support authors of subject submissions, and coordination of inputs that cut across academic departments (see Box 1).

243. It is also clear that, for providers, efficiencies could be made and some barriers may have been due to providers being in the start-up phase. Providers are looking to streamline processes with internal quality assurance and enhancement processes in order to reduce burden:

> "TEF main contacts were asked what the key learnings were from participating in the pilot process. The most common responses were around operationalising the subject-level TEF process (79 per cent), finding ways of embedding subject-level TEF into existing quality assurance and teaching enhancement process (76 per cent), and understanding the performance of own subjects against those in the wider sector (76 per cent). The point about embedding the requirements for subject-level TEF within existing internal quality assurance processes came through strongly in the depth interviews as institutions look for ways to streamline processes and reduce the added burden to staff of subject-level TEF."
>
> TEF subject-level pilot evaluation – Provider perspectives report

> "Another barrier to engagement for academics was the metrics. The quantitative data determined a considerable lack of understanding of the metrics. Insight from depth interviews shows that a lack of understanding was only part of the issue, with some academics expressing concern about the validity of some of the metrics as a measure of teaching quality."
>
> TEF main contact, University
> TEF subject-level pilot evaluation – Provider perspectives report

244. The volume of quantitative information that providers and panel members had to deal with in TEF has increased over time, and was a widespread concern (see paragraph 114). Nonetheless there were also data issues where further improvements would be welcomed, in particular the inclusion of regional benchmarking in the LEO metrics, and the further development of differential attainment data (see paragraphs 114 and 123).

> "The information available to panel members increased in both volume and complexity between the first and the second pilot. The metrics workbook contained a substantial amount of information, from the provider contextual data, new basket of nine core metrics and associated split categories, to new attainment data at provider level and course contextual data at subject level. Panel members recognised that the information available could provide rich and valuable insights, but that it could also be overwhelming and difficult to navigate in real time, as in panel meetings or when asked to be an 'extra reader' on a tricky case."
>
> Main panel chair's report

245. It is important to note also that there was significant logistical complexity to delivering the pilot. Meetings required careful sequencing in order to manage the flow of information between subject and provider level assessment. From a communications perspective there were differing information needs for various panel members (particularly for subject panel chairs and deputies, who needed information for both the provider-level meetings and the subject level meetings). From an events perspective, the physical spaces and agendas required careful co-ordination to manage simultaneous meetings and to cater for panels with differing workloads within the limited time available.

246. To deliver the exercise the OfS used a number of IT systems which had been inherited from previous exercises. This was deliberate, to avoid investing in new systems that might not ultimately be required. However, old systems required significant manual adjustment and identifying work-arounds was time consuming and inefficient.

247. To ensure a better experience for all parties involved, further preparation time to avoid such issues would be vital (see paragraphs 59, 109-110 and Box 1: Provider perspectives ).

248. Engaging students in the process was one area found to be particularly difficult for providers:

> "Around one in three (36 per cent) TEF main contacts said that it had been difficult to engage student representatives, and nearly double this proportion (65 per cent) cited difficulty engaging student contributors."
>
> TEF subject-level pilot evaluation – Provider perspectives report

> "The key barriers to engagement were time, and that students could not see any benefits for themselves in getting involved with the process. Very few said they were ideologically opposed to provider-level TEF in any form (16 per cent), though fewer were opposed to subject-level TEF specifically (just 4 per cent)."
>
> TEF subject-level pilot evaluation – Provider perspectives report

249. For panel members, time pressure was a constant factor, but the overall sense of data-overload stymied their ability to expedite decisions efficiently, undermined confidence and did little to resolve previous concerns around metrics capture. Further time would be required throughout the process for more effective training, more careful calibration and moderation, and to produce high-quality statements of findings (see paragraphs 84-88 and 102-110).

> "The panel found the metrics adequate to interpret, and the systematic nature of the process satisfactory, but noted a number of concerns. These concerns increased the burden on panel members compared with previous years. [...] There was acknowledgment that the TEF exercise remained weighted towards the metric outcomes and that greater emphasis needed to be placed upon the holistic submission."
>
> Social Sciences, and Natural and Built Environment panel report

- **Care would need to be taken to ensure subject-level assessment is fair for FECs and smaller or specialist providers**. There were apparent differences in the organisational capacity and ability of different types of providers to engage in subject-level data and submissions. The pilot outcomes also differed by type of provider. These issues would need to be explored further to ensure subject-level assessment is as fair as possible to all participants.

250. Evidence from providers across both pilots indicated that subject-level TEF would require extensive financial and time inputs. Burden was felt most by FECs and some (but not all) smaller providers. It was notable that the IFF surveys reported that only 23 per cent of academic contributors and only 14 per cent of TEF main contacts agreed with the statement that subject-level TEF, in its current form, would support diversity of provision.[46]

251. To date, the provider-level TEF exercise has shown excellence is spread throughout the sector. The results of the pilot suggest that FECs have a less positive profile of ratings, both in provider and subject-level assessment (see Chart 9 and Chart 10). Part of this pattern is likely to be explained by the impact of missing data and small cohorts, which is more prevalent in these provider types due to the provision they offer, and lower available resources. Care should be taken in extrapolating these results; however, regression analysis undertaken by the OfS also indicated that FECs and providers with fewer than 10 subjects were likely to receive a lower rating, even when accounting for initial hypothesis performance (see Annex G).

252. In reading these findings it is important to note that:

    a. Historically, the highest awards have been found across all provider types in the provider-level exercise. The figures in Table 13 below illustrate this by provider size and region of provider.

    b. The pilot sample was selected to reflect the diversity of the sector and deliberately included higher numbers of FECs and providers with fewer than 10 subjects (see Table 4 to Table 8).

    c. Provider participation will in future be a condition of the regulatory framework: 'Condition B6: The provider must participate in the Teaching Excellence and Student Outcomes Framework'.[47]

[46] See 'TEF subject-level pilot evaluation – Provider perspectives' report, page 5. Available at www.officeforstudents.org.uk/publications/tef-findings-from-the-second-subject-level-pilot-2018-19/.

[47] See: www.officeforstudents.org.uk/advice-and-guidance/regulation/conditions-of-registration/initial-and-general-ongoing-conditions-of-registration/.

### Table 13: Current provider-level ratings, by provider size

| Size (no. of students)[48] | Bronze | Silver | Gold |
|---|---|---|---|
| **Fewer than 500** | 27 | 42 | 10 |
| **500 - 4,999** | 24 | 38 | 27 |
| **5,000 or more** | 9 | 52 | 39 |

253. Allowing for the caveats above, the OfS analysis of the pilot outcomes (see paragraphs 161-218) does indicate different outcomes by provider type in the pilot. The analysis showed that in the pilot:

    a. at provider level, FECs received a lower percentage of Gold ratings (see Chart 9: Final ratings by provider type)

    b. at subject level, large multi-faculty providers received proportionally more Gold ratings and again FECs received the least (see Chart 11: Subject ratings by provider size)

    c. larger percentages of subjects at FECs were 'moved down' from the initial hypothesis and smaller percentages of subjects at FECs were 'moved up' from the initial hypothesis (see Chart 12: Differences between subject ratings and their initial hypotheses by provider type).

254. Issues with data availability also disproportionately impacted certain types of providers. Analysis provided in the 'Findings from the first subject pilot'[49] illustrated how:

    a. there was a neutralising effect, whereby subjects with smaller cohorts and fewer available data sources would 'default' to a Silver initial hypothesis, which could in turn be difficult to assess.

    b. introducing thresholds to exclude subjects with limited data (i.e. by excluding subjects with partial metrics and/or low numbers of students within a subject instance) would result in significant coverage issues for providers who teach across a range of subjects with low student numbers – most typically FECs.

255. Findings from the second pilot confirm that it would be necessary to apply such thresholds for assessment and also that, where providers are just above these thresholds (i.e. with partial data that is deemed under current rules as being assessable), they are likely to receive poorer results (see Chart 15, Chart 16, Chart 17 and Annex H).

256. The OfS examined costs to providers in the first pilot and found that whilst FECs and alternative providers spent less on the pilot in absolute terms, the cost per student was

[48] Size is calculated on the basis of how many undergraduate students (FPE) were actively studying at each provider in 2017-18.

[49] TEF Findings from the first subject pilot, paragraphs 94–100. Available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-findings-from-the-first-subject-pilot-2017-18/.

higher. These results, originally published in October 2018 (Annex D: Provider cost survey)[50], were found to be statistically significant and higher relative costs were clearly identified for these types of provider.

257. Whilst IFF did not systematically capture time and costs for this pilot, academic leads at providers (those responsible for writing subject-level submissions) did provide self-reported information on the time they spent participating in the exercise. This suggests that in many cases FECs and alternative providers were not able to invest as much time in submission development (note that the figures given below refer to time spent by individual contributors but providers would typically have had numerous academic leads. These figures also do not include time spent by, for example, central staff or student contributors).

> "The demands on staff time were considerable. More than half of academic contributors said that they spent between one and two weeks in total on the process (56 per cent); one in four (25 per cent) said that they spent more than two weeks on subject-level TEF."
>
> TEF subject-level pilot evaluation – Provider perspectives report

258. The IFF research also reports that there are capacity issues in some providers, and that some providers may have better developed internal systems:

> "In contrast to universities, which are often larger and more established, learnings for other providers were more focused on collecting a broader range of metrics and establishing a more standardised process across departments going forward."
>
> TEF subject-level pilot evaluation – Provider perspectives report
>
> "While it is a small sample size, it is worth noting that no TEF main contacts within Further Education Colleges or Alternative Providers reported that their academic contributors found it easy to contribute to the TEF subject-level pilot process. Instead, they were most likely to report that academic staff found it 'fairly difficult' to contribute (64 per cent and 60 per cent respectively). A similar, if less extreme, pattern occurred with respect to senior leaders, where those at universities tended to find it easier to contribute as well."
>
> TEF subject-level pilot evaluation – Provider perspectives report

259. These disparities extended into other areas; for example, it was often noted that FECs and alternative providers have different student representation structures, and their students may have other barriers. As one student contributor at an FEC told IFF:

[50] Available at: www.officeforstudents.org.uk/publications/teaching-excellence-and-student-outcomes-framework-findings-from-the-first-subject-pilot-2017-18/.

"Many students are part-time with work and family commitments, so it was difficult for them to find the time to engage with the process."

Student contributor, Further Education College
TEF subject-level pilot evaluation – Provider perspectives report

## Conclusion

260. It is valuable to consider higher education at subject level: there are many subjects that deserve recognition and to be celebrated for the excellent experience and outcomes their students receive. But we also know that the quality of teaching provision varies significantly within universities and colleges. Understanding and addressing this variation has the potential to drive excellence in learning and teaching, and to provide more useful information to students.

261. This second year of piloting has been incredibly valuable. We have found that the model for the second pilot was an improvement on the previous year's pilot models and the pilot demonstrated that scrutinising subject level evidence could help drive improvements in higher education. The panels developed substantial expertise and the processes relating to preparation, training and assessment were refined.

262. Whilst there is scope to extend the evidence and assessments within TEF to subject level, the pilots and further metrics analysis show that we do not yet have a model for producing robust subject-level ratings across the sector that is ready to implement. There are significant limitations in the data at subject level, and other issues that would require further development before moving to a scheme that systematically rates all subjects.

263. The OfS will carefully consider the findings from this report, alongside the recommendations from the independent review and any guidance provided by the Secretary of State, and we will consult on a proposed approach in 2021.