TEF at subject-level: summary of findings from the 2017-18 pilot

Key Points

- Ratings were successfully generated at provider and subject-level in both models using the existing provider-level framework. Both models feature design elements that were intended to reduce burden but ultimately added to the complexity of the exercise and did not produce robust ratings for all subjects. Panel members and providers expressed support for a more comprehensive model, drawing on the best elements of Model A and Model B.
- The method of assessment based on combining metrics and submissions worked well but there were limitations in using the data at subject-level. There was a tendency for metrics to 'default' to a Silver initial hypothesis and panels found they were unable to reach judgements, or were not confident in the judgements they made, where there was a sparsity of data arising from small subject cohort sizes and missing data sources.
- Grouping subjects at the second level of the Common Aggregation Hierarchy (CAH2) level provided the right level of aggregation. However, some refinements to the classification could be made to better reflect diversity of provision. The teaching intensity measure was a significant burden to providers and was not found useful by panel members.

How the pilot was conducted

- Two models were tested in the first subject pilot. For both, assessment was based on the same criteria and evidence used in provider-level TEF. Importantly, TEF ratings were produced for each subject and for the college or university as a whole.
- Model A is a 'by exception' model. Subjects were identified as 'exceptions' where the metrics perform differently to the provider-level metrics. These exception subjects were fully assessed and received a rating of Bronze, Silver or Gold. Non-exception subjects were not assessed, and received the same rating as the final provider rating.
- Model B is a 'bottom-up' model. All subjects were assessed as part of a 'subject group' submission but with separate metrics for each subject[1], and each subject received a TEF rating of Bronze, Silver or Gold. The subject ratings then fed into the provider-level assessment through a subject-based initial hypothesis[2]. In Model B provider-level submissions had shorter page lengths and focused on three of the 10 TEF criteria.

[1] The 7 subject groups used for assessment under Model B comprise the 35 CAH2 subjects placed into categories with similar subjects e.g. the 'Business & Law' subject group comprised the Business and Management and Law CAH2 subjects. The 'Humanities' subject group comprised eight different CAH2 subjects e.g. Communications and Media Studies, English Studies, etc.

[2] The subject-based initial hypothesis for Model B is calculated by weighting the final subject ratings for a provider by the number of students studying each subject. This result is then considered alongside the provider metrics and the provider submission to determine the provider's overall rating.

Who took part in the pilot?

- A total of 50 colleges, universities and other types of higher education provider took part in the first subject pilot. They were selected to reflect the diversity of providers across the UK.

The panels

- Over 140 panel members carried out the pilot assessment process. They were selected for their standing in the higher education sector, expertise, and commitment to excellence in teaching.
- The Main Panel decided ratings at provider-level. The seven subject panels decided ratings at subject-level. Each panel consisted of student representatives, academics, and representatives from employers, widening participation experts and professional, statutory and regulatory bodies (PSRBs).

Timings

- The pilot consisted of three broad phases: preparation, submissions and assessment. Providers participating in the pilot received subject-level metrics at the beginning of December 2017 and had a three-month window to complete their submissions.
- Assessment took place between March and May 2018. Model A and Model B assessment processes were carried out separately but conducted by the same panel members to inform a comparison of the models.

Main Panel Chair's report

- Members across the panels felt that robust ratings were generated, with an important caveat at subject-level: where metrics data was sparse, for whatever reason, panels had less evidence to inform their judgement and felt less confident in awarding a rating to these subjects.
- The opportunity to carefully agree 'No rating' for a minority of subjects, where panels felt evidence was missing across both metrics and submission, increased confidence in the process and the robustness of ratings awarded.
- The fullness of assessment in Model B made it a more comprehensive and reliable source of information for applicants. While the subject group submission and the subject-based initial hypothesis were weaknesses in Model B's design, overall panel members saw more opportunities for tweaks to improve this mode.
- Model A was felt to be lacking in its provision of information to applicants and potentially misleading, with concerns around the unknown element of non-exception subjects that would inherit the provider-level rating with little interrogation.
- One of the key benefits of TEF, and the future benefits of subject-level TEF, is increasing the focus on teaching enhancement. Panels felt that this would not occur across the board if only a sample of subjects were fully assessed. Non-exception subjects would not have an associated submission, making it difficult for students to make fully informed choices.

Quality of the evidence

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

The assessment process

- The Gold, Silver and Bronze rating descriptors did not seem entirely appropriate for subject-level ratings and could be tweaked at both subject and provider-level. There was support within the Main Panel for keeping subject-level and provider-level submissions, but decoupling them, potentially extending to using different criteria to judge them.
- Panels found that there tended to be 'stickiness' of the metrics in assessments, and a better balance between metrics and submissions is needed to promote the holistic judgement of providers and their subjects. This was particularly the case in Model B, where a lower proportion of subjects were moved up than in Model A, and the panel felt that this was due to the methodology, rather than reflecting the genuine spread of excellence.
- Generally, level 2 of the Common Aggregation Hierarchy (CAH2) subject classification worked for subject-level assessments, although most panels identified one subject in their subject group which was a 'mixed bag' of courses or did not align entirely naturally with the rest of the group.
- Evidence of impact was the key indication of good quality submissions at both provider and subject-level. The best submissions addressed both strengths and weaknesses in their data, showed an understanding of their mission and their students, and were able to identify where interventions had resulted in a clear positive change.

Potential impacts

- Further work is needed on how subject TEF will inform student choice - both current models could result in difficulties in the way subject ratings would be presented to potential applicants, and getting this right is critical to the integrity of TEF.
- The messaging around subjects that do not have sufficient data for assessment must be carefully considered. How subject-level information is communicated becomes vital and should be a consideration in the second subject pilot (2018-19).
- Another consideration at subject-level was the lack of data for new courses, or the sometimes poorer data in areas where changes in curriculum had been introduced. Not giving a rating, or giving a Bronze rating, to a course that was new or undergoing significant change, did not feel to some panel members like a fair outcome.
- The costing of the exercise, particularly in terms of a scaled-up panel assessment process, needs to be thoroughly assessed ahead of full implementation. A full sector subject-level TEF would require a panel greater in size by an order of magnitude and would require a comprehensive recruitment exercise.

Synthesis of findings

- Across both models we received and considered 727 subject and provider-level cases. The exercise demonstrated that subject-level assessments and decisions about the ratings can be successfully made based on the framework currently applied in provider-level TEF.

How the models generated subject-level ratings

Model A

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

- However, there were concerns about the key design feature of Model A: its exception-based approach to subject-level assessment. Respondents commented on the complexity of the Model A process for generating exceptions, and there was a mixed view among providers on whether the process identified subjects that were known to have a stronger or weaker performance than the provider overall.
- From the outset a key question surrounding Model A was the validity of its central premise: that the provider-level rating produced by this model would reflect teaching quality and student outcomes in most parts of the provider's subject provision. If this premise is invalid, there is the potential for 'non-exception' subjects to automatically inherit a provider rating that does not reflect the actual subject-level provision.
- Following the completion of the Model A provider-level assessments, we designed and conducted an exercise in which we asked a subset of the Main Panel to review an allocation of providers' entire metrics (i.e. the full set of provider and subject-level metrics for each provider).
- Reviewers reported back on how reasonable it would be for each subject within a provider to inherit the final provider-level rating rather than being fully assessed by the relevant subject panel. Each provider was reviewed twice, and we then examined the level of consensus between the two reviewers.
- Across these combinations we found low levels of consensus that it would be reasonable for a subject to inherit the provider rating. For example, when the final provider rating remained consistent with the provider initial hypothesis, there was consensus for only 52% of non-exceptions that it was reasonable for those subjects to inherit the provider rating without full assessment at subject-level.
- This evidence is further supported by the subject ratings generated in Model B and their relationship with the final provider-level rating. Approximately 30% of subjects that would have been classified as non-exceptions in Model A received a rating that was different to the final provider-level rating.
- That is, using Model B ratings to simulate Model A outcomes also indicates that a significant proportion of provider-inherited ratings for non-exception subjects would not represent provision as accurately as full assessment at subject-level.

Model B

- The key advantage providers and panel members identified in Model B was the potential to address all subjects offered by the provider. However, providers and panel members broadly agreed that the current format of provider and subject-level submissions in Model B did not allow for the totality of an institution's provision to be represented or assessed.
- The subject-group submission format was also challenged by providers in the cost survey, where written comments suggested that workload actually increased based on the amount of editing and condensing required on the submission.
- Further concerns were raised by students – who saw less evidence of student engagement in Model B submissions – and by widening participation experts, who similarly reported that widening participation concerns were less visible in Model B submissions.
- There was consensus across the panels that Model A subject submissions, where they were made, provided a foundation for better subject-level assessment than the subject-group submission format in Model B. This is reflected in the fact that they appear to have played a greater role in decision-making than in Model B – a higher proportion of assessed subjects received ratings that were higher than the initial hypothesis in Model A than in Model B.

- Model B led to more complex and lengthy discussions, as was expected given the need to consider subject-level ratings through the subject-based initial hypothesis (SBIH), and a more limited provider-level submission.
- Overall, the SBIH was found to cause anchoring and to compound issues of 'metrics capture' (which refers to panel members being reluctant to move their holistic judgement away from the position indicated by the metrics alone) and 'silverness' (which refers to the higher rate of Silver initial hypotheses observed at subject-level, compared with provider-level).

A more comprehensive model

- At OfS student focus groups, there was a consensus among participants that Model A produces outcomes which are less useful to applicants. For example, a final provider-level rating may well differ from its initial hypothesis, and students would be unaware of how a non-exception subject 'sits' in the context of its inherited provider-level rating.
- The focus groups felt strongly that a model which fully assesses each subject would produce a more meaningful and accurate set of ratings for applicants and students. This view was supported by student panel members and student representatives.
- Crucially, 10 out of 12 providers that participated in both models agreed or strongly agreed that the least time-consuming approach to subject-level TEF would be one that required a submission from each subject (e.g. five pages) and a provider-level submission (e.g. 15 pages across all criteria).

Rating structure and descriptors at subject-level

- Subject panels reflected dissatisfaction with using the existing rating structure at subject-level. Subject panels found that the ratings Bronze, Silver and Gold applied to too large a range of performance at subject-level. In particular the gap between 'low' and 'high' Silvers was problematic, and borderline ratings required significant debate to resolve.
- These issues were compounded by the associated ratings descriptors, which were not tailored enough at subject-level. The subject panels recommended the refinement of these descriptors.

Assessability of subjects

- Reportability: Some subjects do not have a full set of reportable metrics. This may be due to very small cohort sizes, new subjects that do not yet have reportable data, particularly employment metrics, or variations in survey response rates.
- Assessability: Some subjects have very low cohort numbers and therefore confidence in the statistical reliability of the data is low. Such cases generate few or no flags, and therefore most cases are Silver 'by default'.
- These findings highlight the need to develop explicit criteria for determining whether a subject should be 'in scope' for subject-level assessment, and additional processes would need to be built into subject-level TEF to confirm which subjects at each provider are in scope for assessment.
- All panels independently came to the view that a cohort size threshold for assessment would be necessary to enable the assessments to be meaningfully informed by metrics. In addition to finding some subjects with insufficient data to inform assessments, panels were concerned more generally by the number of subjects that 'defaulted' to Silver in the initial hypothesis at step 1a. This 'silverness' at subject-level is caused by the neutralising effect of fewer flags when there are smaller cohorts and fewer reportable data sources.
- The need for thresholds raises questions about how subjects falling below the thresholds should feature in subject-level TEF, and the impact they would have on the coverage of subject TEF ratings.

The panels agreed that 'No rating' was unhelpful terminology and an unsatisfactory solution, and that it would be appropriate to consider different options.

- For example, it may be appropriate to consider provisional (or similar) awards for new courses or subjects with small cohorts, but no awards for subjects that do not offer whole courses or that have been completely taught out.

Accreditation

- In general the panels did not feel that a compulsory declaration of accreditation should be a requirement. As the significance of accreditation varies widely across subject areas, developing a standardised approach was not viewed as useful.

Interdisciplinarity

- Providers and panel members found the articulation and interpretation of narratives for the three broad interdisciplinary subject categories challenging in a number of cases. For some providers, the diversity of courses mapped to these categories meant it was difficult to produce a submission that fully addressed very different student experiences.

Teaching intensity

- The majority of providers found it too difficult or too resource-intensive (or both) to accurately and robustly capture teaching intensity information in its current form. The survey received only 4,880 full responses out of approximately 113,000 responses (a 4.3% response rate).
- Providers found the generated metrics difficult to interpret and their use in provider submissions was therefore limited – more than half of providers did not refer to teaching intensity data in their submissions at all.
- Similarly, no panel found it possible to meaningfully interpret teaching intensity data in its assessments. Therefore it was felt that no judgements of quality surrounding teaching intensity could be made, and panels reported that it did not influence the final ratings.
- If teaching intensity was a mandatory element of subject TEF, it might represent a disproportionate cost compared with the rest of the exercise. Estimates for both Model A and Model B indicate that returning teaching intensity data for all subjects on average comprised approximately 50% of the total cost of participating in the pilot. This indicates that teaching intensity does not provide value for money.

Grade inflation

- In the pilot, grade inflation information was available for the provider-level assessment of providers with degree awarding powers. Panel members observed that some of the provider narratives surrounding grade inflation were of interest, but ultimately found it difficult to resolve the fact that the intended measure of grade inflation works in tension to one of TEF's key purposes: to encourage enhancement in teaching and outcomes for all students, of which attainment is an aspect.
- Main Panel members reported that grade inflation data had little or no impact on their holistic assessment. It was suggested that grade inflation should be dealt with as a standards issue through regulation and other sector-led initiatives. The panels were also concerned that the incorporation of grade inflation data into assessment sent contradictory messages in relation to the pressing need to close gaps in attainment between students from advantaged and disadvantaged groups.