Office for National Statistics

Methodology Advisory Service

Evaluation of the Statistical Elements of the Teaching Excellence and Student Outcomes Framework

A report by the Methodology Advisory Service of the Office for National Statistics

Gareth James
Daniel Ayoubkhani
Jeff Ralph
Gentiana D. Roarson

July 2019

Contents

List of tables

List of figures

List of acronyms

BME: Black and minority ethnic

CAH: Common Aggregation Hierarchy

DfE: Department for Education

DLHE: Destination of Leavers from Higher Education

EU: European Union

GO: Graduate Outcomes

GSS: Government Statistical Service

HECoS: Higher Education Classification of Subjects

HESA: Higher Education Statistics Agency

IMD: Index of Multiple Deprivation

JACS: Joint Academic Coding System

LEO: Longitudinal Education Outcomes

NAO: National Audit Office

NSS: National Student Survey

OfS: Office for Students

ONS: Office for National Statistics

PG-UG: Postgraduate-undergraduate

POLAR: Participation of Local Areas

RSS: Royal Statistical Society

SOC: Standard Occupational Classification

TEF: Teaching Excellence and Student Outcomes Framework

UCAS: Universities and Colleges Admissions Service

Executive summary

We provide a relatively short, and mainly non-technical overview of the report and main findings here. For brevity, we have necessarily omitted some aspects from this summary, and focus on the areas we regard as being most salient.

1. Introduction

This report evaluates the statistical elements of the Teaching Excellence and Student Outcomes Framework (TEF). It has been undertaken by a team of experts at the Office for National Statistics (ONS) comprising of methodologists, statisticians and social researchers. This evaluation was commissioned by Dame Shirley Pearce, through the Department for Education (DfE) and under an agreed scope, as part of her wider review of TEF. The areas identified as priorities for our work were, broadly, the statistical methods themselves, and the communication of statistical concepts. We note that, naturally, TEF has various policy aims, which should be supported by data and statistics. We have not explicitly assessed the extent to which the statistics do that, rather we have appraised them against statistical and methodological good practice.

Our evaluation reflects an intense period of learning and investigation by our team over a short period of time. It represents a 'fresh and independent look' at TEF methods, as the team's involvement with TEF has, mainly, been quite limited to date. Most of our expertise lies in methods related to the broad range of official statistics. The review has been conducted independently by scrutiny of publicly available documentation and analysis of TEF data supplied to us by the Office for Students (OfS), together with interviews with TEF assessors and Panel members. We have sought clarification only where necessary and to aid our understanding; any errors in this report remain our own. We have also been mindful of the views of others about TEF, such as those of the Royal Statistical Society (RSS).

This report contains our comments on the data, methods and statistical processes of TEF, as considered in comparison with statistical good practice; it also contains analysis of the impact the current methods have on the TEF assessments and outcomes. We have made 33 formal recommendations – categorised as high, medium or low priority – for improvements to methods, the communication of statistical topics, and some wider issues for further consideration and research. Where time has permitted, we have suggested options to start exploring our recommendations more fully, but our remit here does not include that developmental work; some additional work will be required for some of our recommendations. Additionally, we also note some other possible points for consideration, but do not always make a formal recommendation. Our thanks go to Dame Shirley Pearce, to colleagues at DfE and OfS, and to the TEF assessors who have been patient with us and helped with our inquiries.

2. Overview of methods

TEF is still relatively new, now being in Year Four (academic year 2018/19). Its methods have changed over time, as TEF has been developed, and we have concentrated our evaluation on the methods currently used. That includes the method used for assessing providers in Year Three and (continuing almost unchanged) in Year Four, and a new method that is currently being piloted (also in Year Four) for both providers and subjects (within providers). We refer to these as the 2017 and 2018 methods respectively.

A TEF assessment is a staged process. The final outcomes are ratings (Gold, Silver or Bronze), proposed and agreed by TEF assessors and a TEF Panel, based on a holistic consideration comprising: a formulaic 'starting point' summary based on quantitative data; a subjective evaluation of other statistical information; and a qualitative assessment of a provider-supplied written submission.

The first stage of the assessment process is called Step 1a, and that has been the focus of our consideration of methods. Six core metrics (nine under the new, 2018 method), covering a range of topics considered to be associated with teaching excellence or student outcomes for a higher education provider, and a variety of survey and administrative data sources, are combined into the single 'starting point' (Gold, Silver or Bronze) via a combination of rules, flags and pre-determined weights.

During subsequent stages of the assessment (Steps 1b and 2), the assessors consider other information on the core metrics (for example, splits that relate to particular student sub-populations), contextual data (other information about the provider and its students) and a 15-page, written provider submission. The TEF Panel and assessors then make a holistic judgement as the final step (Step 3).

At the heart of the Step 1a calculations for any provider are the differences (denoted *d*) between the metrics' indicator values (in each case, the indicator is the percentage of something positive or desirable, such as students who continue in their studies or rate the teaching as positive) and their respective benchmarks (an average or expected value of the indicator nationally, adjusted to take account of the demographic mix of students at the provider). In general terms, it is intended that a large and positive value of *d* be interpreted as the provider being above expectations in terms of teaching quality, the learning environment or student outcomes; such metrics are flagged with a '+' or '++' sign. Those metrics with a large and negative difference may be regarded as indicating the provider is below expectations (and are marked '-' or '--'). Whether or not a metric is flagged is determined by the size of the difference, *d* (values near zero are not regarded as being materially important), and the value of *d* must be also be significantly different from zero in a statistical sense (so we can say the result was unlikely to have occurred by chance alone) – that aspect is determined by the calculation of a z-score (calculated as *d* divided by the estimated standard deviation of *d*).

Each of the six or nine core metrics has an associated weight; these are 0.5, 1.0 or 2.0, and if a flag is realised (separately for positive and negative flags), then the flag value of the metric is assigned the respective weight. A set of rules then combines the total flag values of positively and negatively flagged metrics, so as to determine the provider's Step 1a starting-point category.

3. Evaluation of Step 1a

Our critique of the TEF methods begins in this section. We comment first on the reference period(s) of the metrics. The core metrics combined in any one TEF year relate to student data from different periods: the various data sources have different lags or timeliness associated with their reporting, and most are pooled over a three-year period to improve stability. Clearer descriptions of target populations and assumptions made would benefit TEF.

Statistical uncertainty is present in the data, in particular because non-response exists in the census-(survey)-based metrics, and more generally because a provider's students are assumed to be just one realisation of some much larger 'super-population'. Monitoring of potential non-response bias and assumptions of super-populations could be made clearer. Other than in the flag creation process, the statistical uncertainty present isn't reflected explicitly in the Step 1a calculations and it does not feature in the later presentation of the data; there is no indication of statistical confidence

in the Step 1a starting point, for example, and we think that communication about this topic should be improved. Further, better guidance about making multiple comparisons (hypothesis tests) should be provided, so users (potential students, TEF Panel members and providers, for example) can more appropriately interpret multiple TEF outputs, in which the risk of flagging insignificant results by chance increases with the number of statistical tests made.

Benchmarking is a central aspect of the calculations. Its intentions are good, and it is important to adjust for different mixes of students at different providers. The benchmarking method used has been developed diligently, has a theoretical grounding, and has a design-based formulation that is arguably more transparent to lay users than an equivalent model-based approach. We suggest that guidance around how the difference, *d*, should (and should not) be interpreted could be improved to the benefit of prospective students. More specifically, benchmarking takes account of factors included in the definition of the benchmarking groups, but obviously does not, and cannot, take account of other differences that exist and are not related to teaching quality, the learning environment or student outcomes and learning gain. Thus that difference, *d*, contains both what TEF is seeking to measure (in broad terms, the value-added by a provider), and some amount of other differences that exist between providers that are unmeasurable or unobserved (ones not included in benchmark definitions), as well as some random noise. We make other recommendations about benchmarking and related aspects too, but these are about the processes themselves, and we consider them to be mainly of lower priority as they address the detail of the current process and the communication of it.

We also consider factors that ideally would be used as benchmark factors but are difficult to measure or only available at provider-level rather than student-level. Examples include student expectations (of a provider) or locality-based factors; these are outside the control of the provider but may have a bearing on the observed metric values. If we were to assume there exist fundamental differences of this sort between providers, and that providers could be grouped in some meaningful and accepted way, then the option to carry out TEF assessments separately by group would have statistical benefit. Some mitigation against not doing so exists, however, if the missing factors are correlated with existing benchmark variables, and maybe also through the TEF assessors' and Panel's holistic view. We recommend research is required to try to establish if suitable groupings might exist, and further consideration as to whether and how TEF assessment by group would be practical to implement.

The descriptive analysis included in Section 3 includes an assessment of the correlation between metrics, and notes a positive association between some of the pairs of metrics (in particular those derived from the National Student Survey (NSS)). That is to say that, based on observation alone, some of the metrics appear to be capturing some of the same information as other metrics; that association led to a previous reduction in the weights of the NSS metrics. We provide a comparison of the weights between the 2017 and 2018 methods. Additionally, we suggest a tool could be developed to allow users, prospective students for example, to input their own weights to reflect what they regard as the relative importance of the metrics; of course, that could only be taken as far as the formulaic Step 1a starting-point calculation.

We next consider the impact of the binary nature of the flags – for each metric, a (positive or negative) flag is either realised or not, and thus collects all or nothing of the available weight – on the Step 1a process. We note that it is possible for two almost identical providers to receive different starting-point outcomes under this approach, and in such a way that the assessors may be less likely to notice or consider either for a change in award. As such, we strongly recommend that a different approach is taken in Step 1a, and probably one where some proportion of the weight can be realised

for differences, *d*, that are found close to the materiality and statistical significance thresholds (absolute values of *d* in excess of 2 percentage points, and of *z* in excess of 1.96). In Annex B, we make tentative suggestions for how this may be developed and implemented.

We consider missing metrics – for example, those that for various reasons are non-reportable – and comment on how they are handled during the process and the effect that this might have. We think this is an aspect of the process that is not well documented, and recommend that this is changed, as well as considering the effect of current practices. We also consider the mode of study (full-time or part-time) and note that only the majority mode is usually included in the Step 1a calculation; we think that there should be scope for including both in a formulaic way.

Finally, in this section, we comment on the principle of combining the various metrics, as is done in Step 1a, in which the six or nine reasonably diverse metrics are combined into a single, overall measure. We don't think there is a right answer as to whether this should be done or not, but set out some considerations both for and against, and advise that these are considered.

Overall, there is scope for various changes and improvements to be made across the Step 1a processes and methods, which will require further development, and should be undertaken in a holistic way.

4. Possible adaptations to the Step 1a process

In this section we start to consider how the Step 1a process could be changed fundamentally from the current approach. We consider the use of a net total value, which we define as the difference between the total of the positively flagged metrics and that of the negatively flagged metrics. Such a measure seems the natural and intuitive calculation, but that and the current starting-point diagrams are quite inconsistent: Gold and Silver starting-point regions, for example, overlap in terms of their net total value: a provider could have a lower net value but get a higher starting point than another provider under this approach, and, again, it seems to be the borderline cases that have the potential to lead to inconsistent and undesirable outcomes. We also note how the current approach seems to penalise negative flags much more harshly than positive flags are rewarded, and ask if this was intentional.

Given we have seen how relatively similar providers can be assigned quite different starting points, and that statistical uncertainty is not captured at this stage, we wonder whether alternative forms of presentation would be useful and workable. Options could include presenting the net total value as a mark on a scale of, say, 0 to 100, with graduated shading used to give an indication of the colour (Gold, Silver or Bronze) without a specified category being assigned.

In addition, it would be relatively simple to develop an analogous but not benchmarked, combined indicator for the same metrics using data already published in TEF workbooks; it should also be possible to develop confidence intervals for such a measure. It may be of use to TEF assessors, and potential students, to see both benchmarked and non-benchmarked indicators presented side-by-side, giving a more transparent view of the effect of benchmarking and allowing more scope to investigate differences and similarities.

We have carried out a number of sensitivity analyses on the published TEF Year Three metrics data based on options to change metric weights, change flag thresholds, and weight together metrics for full- and part-time students, and we report our findings in this section.

Our main recommendation is that the Step 1a process should be developed further, taking into account the various issues noted and suggestions made to make the starting-point outcomes more consistent. Dependent on the developments undertaken, the Step 1a outcomes could potentially be presented in a such a way that includes measures of uncertainty.

5. Contextual data

Most comment in this section is confined to principles around the methods for the 'high and low absolute values' reported. We note some implications of the methods (mainly that smaller providers are less likely to be marked as significantly high or low), and suggest that improvements are made to some of the documentation.

We also report descriptive statistics on the metric-split categories, noting that small sample sizes can be prevalent, which raises questions about the validity of any statistical inferences based upon them. Additional guidance for TEF users about the implications of small sample sizes would be useful.

6. Statistical infrastructure

This section considers the topics of harmonisation and classifications. The Government Statistical Service (GSS) maintains a number of harmonised topic standards for questions, question-wording, answer categories, and output aggregation-group hierarchies that have been developed and tested so that they can be applied consistently across official statistics. We have compared these GSS-wide standards to topics found on the Higher Education Statistics Agency (HESA) student record, which guides TEF metric splits and contextual information, for example. Overall, we found many of the definitions to be consistent, or largely consistent, but have noted instances where there is a divergence and we think a change would be beneficial.

On classifications, we make comments on TEF's use of the Standard Occupational Classification (SOC) and the Common Aggregation Hierarchy (CAH) [of academic subjects, though that is not immediately obvious from its name].

On the former, we note that SOC is used for graduates' occupations, which is reassuring (however, we have not examined the quality of the coding process). The use of SOC in defining highly-skilled employment in the core metrics also seems reasonable, but it is worth taking into account that future revisions to SOC will cause discontinuities, and we advise that plans are put in place to manage this.

In terms of CAH, we first note an imminent discontinuity when the subject-level codes change from the Joint Academic Coding System (JACS) to the Higher Education Classification of Subjects (HECoS) in the autumn of 2019.

The use of CAH allows consistency across statistics in the higher education sector, and the CAH Level 2 codes are used to define the subjects in the subject-level TEF pilot. We note some observations regarding this for TEF, which may affect the usefulness of the subject-level statistics. First, there is a balance to be struck between having too broad a grouping (which will provide a more robust measure, but will inevitably result in subjects that are of interest to a potential student being 'bundled in' with other subjects that are not of interest), and too fine a grouping (on which statistical inferences might be less valid): there's no right answer to that. We also note, via an example, of how the CAH Level 2 subject groupings might cut across a provider's organisational structure, and thus one TEF rating might reflect the input of two or more organisational structures' different teaching practices. We are not convinced such an output would be of particular use to either potential

students or providers looking for specific information to help bring about improvements. Unfortunately, we have no particular recommendations to offer on this, other than to consider the usefulness and usability of the subject-level assessments defined according to this classification. Providers and potential students might benefit more by consulting the more-detailed CAH Level 3 results of NSS, for example, albeit with caution when making inferences.

7. Communication

Throughout the report we have noted instances where communication and documentation could be made more explicit and have identified gaps that could be filled; we note that we found it difficult to assess during our work whether we had all the information and data available about TEF and in its most up-to-date version. TEF is a complicated subject and having a more consolidated repository of metadata would be useful.

More generally, on clarity and transparency, we have described five broad user groups: the TEF assessors and Panel; higher education providers; the media and specialist press; students and potential students; and the Universities and Colleges Admissions Service (UCAS). We have set out the different uses we think these groups would make of the TEF information. We recommend that this is developed further, and users are consulted on their experience of using the existing outputs and asked about improvements that would help them in their use.

Using information from investigations into the data and methods and their quality, described in earlier sections, we have looked at how that is fed through into the communication of the TEF awards and associated data. The implication of the inherently complex calculations and statistical uncertainty could be better explained.

We have looked at the public web-based delivery of the TEF information and make recommendations for changes that should help users steer through this.

8. Update on 2016 review

In 2016, ONS conducted a review of the data sources in TEF, and made six recommendations. We briefly consider developments in TEF since then, including the development of the Graduate Outcomes (GO) survey as a replacement for the Destination of Leavers from Higher Education (DLHE) survey. We consider that three of the recommendations are now complete, and three (those about under- and over-coverage, non-response bias, and the benchmarking process) are ongoing.

9. Appropriateness of TEF metrics

We have considered the nine core metrics of the 2018 method against eight criteria set out by the National Audit Office (NAO) in their ‘Choosing the Right FABRIC’ document (NAO, 2001). Although most metrics compared well with most criteria, we consider that all metrics are at some risk when compared with ‘Avoiding perverse incentives’ (essentially gaming), and that some metrics (‘Continuation’, and those associated with post-education employment) may be less attributable to the provider. In addition, there are other aspects of the ‘Highly skilled employment or further study’ metric that could be improved. We have not assessed how well any of the existing metrics functions as a proxy for what TEF seeks to measure, nor have we tried to identify other potential data sources for TEF.

10. Use of statistical information by the TEF assessors and Panel

On the edge of the scope of our evaluation of the statistical elements of TEF, we have considered what happens to the data (Step 1a and contextual information) in the Step 1b and Step 2 considerations by TEF assessors and the TEF Panel.

Interviews with four assessors suggest that a robust process is in place for provider-level assessments, but the same was not true for the subject-level assessments being piloted. Useful feedback emerged about how statistical information is used and regarded, especially by assessors with less statistical knowledge. There seems to be a difference in how some provider types are treated; in some cases, supporting evidence is sought in the submissions, whereas in others the submission is used to fill gaps in the metrics data. It is clear that the process is regarded as holistic, and that the Step 1a outcomes are considered as just a starting point for the deliberations. Assessors were asked if further statistical information would be useful, but none was specifically identified.

We have also conducted an analysis of the change in award observed between the Step 1a starting point and the final (post-Panel) outcome. Some patterns emerged from this analysis, albeit based only on a small number of providers, suggesting that some types of providers (for example, higher education institutions) were more likely to have their Step 1a starting point upgraded than others (for example, alternative providers and further education colleges).

11. Conclusions

Our main conclusions are that the TEF methods have been developed with a lot of care, and it is commendable that we now have an assessment of a very complicated and difficult-to-measure concept. However, the statistical elements of the current process have the potential to produce inconsistent results, and we think these can be improved. As already noted, there are elements of the communication of TEF, its documentation and guidance on usage, that could be enhanced too.

Evidence for a number of aspects of our investigation point towards different types of provider being treated differently in some way, and that Gold for a provider of one type currently has a different meaning to Gold for a provider of a different type, though that is not explicitly stated. As already discussed in this summary, we recommend that that further, substantial research is carried out to determine the details and feasibility of grouping providers by type for TEF assessments. Likewise, different pieces of evidence – for example sample size considerations, the subject classification groupings and comments from TEF assessors – suggest that the subject-level assessments are less robust than the provider-level assessments, which could limit their usefulness, although better communication and other, more general improvements to TEF could provide some mitigation; however we suggest that the subject-level TEF pilot, and lessons learned from it, are carefully reviewed.

12. Closing remarks

This report is a record of the investigations and analyses we have undertaken regarding the statistical aspects of TEF. Given the relatively short time available, the evaluation has necessarily been brief; nonetheless, we hope the recommendations we make will assist in forming future development plans for TEF and suggest that any changes to methods are considered in a holistic way.

Summary of recommendations

The list below is a summary of the recommendations made in this evaluation, and is presented in the order in which they appear in the report. We have categorised each recommendation as relating to:

- technical issues: specific areas in which changes may be required in formulas, processes, definitions, question-wording, and likewise; any new developments would then need to be communicated clearly
- communication: areas in which the communication of TEF statistical topics should be improved
- wider issues: those we consider to be broader than a single, technical aspect and which would likely require careful consideration and consultation.

Naturally, there are other categorisations or groupings of recommendations that could have been used (for example, those relating to benchmarking aspects, or the Step 1a process). We have simply presented the recommendations in the order they occur in the report; in terms of next steps, however, it may be useful to consider recommendations on similar topics together.

We have also rated each recommendation according to priority (high, medium or low) in terms of the impact we believe the suggested improvement would have, or conversely, the seriousness of not addressing the issue.

Finally, we acknowledge that whilst some of our recommendations can be implemented with relative ease, others will undoubtedly require considerable effort and time, particularly those necessitating further research or the elicitation of expert opinion. We do not believe it is within our remit to quantify the resource intensity associated with each recommendation, so we make no attempt to do so here; rather, we view this as being an important factor that will need to be considered and weighed up against the priority ranking of the recommendations when they are acted upon.

Rec. No.	Section No.	Recommendation	Categorisation	Priority
1	3.1	Time series analysis, including an assessment of stability, of TEF-input core-metric indicator series should be conducted and made available on an ongoing basis.	Technical	Low
2	3.2.1	Documentation describing the NSS confidence-interval calculations, and the assumptions upon which they are based, should be updated and made clearer.	Technical and Communications	Low
3	3.2.1	Monitor and report on potential non-response bias in the NSS (and other sources) on a regular basis, introducing appropriate non-response weighting or calibration if required.	Technical and Communications	High
4	3.3.2	Further consider Studentisation (the removal of the contribution of each provider from its own benchmark) in the benchmarking process, including its implications for other parts of the process and the possible impact on robustness due to small sample sizes.	Technical	Medium

5	3.3.2	Consider refining the benchmark model building process, for example by investigating variable reduction techniques and cross-validation; and summarise this process in the TEF user guidance, including appropriate model outputs and diagnostics indicative of goodness-of-fit (such as pseudo-R^2).	Communications	Low
6	3.3.3	In the published TEF documentation, make it clearer that the TEF outputs cannot solely be attributed to teaching quality, learning environment, and student outcomes / learning gain, because the benchmarking process does not take into account all confounding variables (those that would meet the TEF benchmarking principles but have not been included as benchmarking factors because they are unobserved or unmeasurable).	Communications	Medium
7	3.3.4	Research further the grouping of providers by type in the TEF assessment process to account for factors that cannot easily be included in the benchmarking process with the aim of improving comparability further. If the research outcomes suggest groupings that are viable, carefully consider the options, implications and practicality of implementation.	Wider	High
8	3.4.2	A fuller description of the target population, and any assumptions made about it, are made more explicit in the TEF user guidance.	Communications	Medium
9	3.4.4	Make more explicit in the publicly available TEF documentation details of assumptions made in the calculations of z-scores and their standard deviations.	Communications	Medium
10	3.4.4	Calibration of the TEF flagging system should be reviewed periodically, on an ongoing basis, and corrective action taken where necessary.	Technical	Medium
11	3.4.6	Improve communication on statistical uncertainty. For example, publish plots of TEF-metric differences and their confidence intervals by provider in rank order of the differences; the plots would clearly show which confidence intervals include zero, and which differences have absolute values that exceed thresholds considered to be meaningful.	Communications	High

12	3.4.6	Appropriate guidance on making multiple comparisons should be drafted and made prominent in the user guidance and with TEF outputs. The recommended plots of differences and confidence intervals could also accommodate this, with further extensions to the intervals' widths for multiple-comparison purposes also shown.	Communications	High
13	3.6	To improve transparency, TEF should adopt use of relative weights for the core metrics, rather than absolute weights.	Communications	Low
14	3.6	Convene an expert panel to decide on the metrics' weights (if not done already), and clearly communicate the principles, decisions, and the rationale for them.	Technical and Communications	Medium
15	3.6	Consider developing a 'personal TEF calculator': a tool that allows users to input their own metric weights for the Step 1a calculation.	Communications	Medium
16	3.7	Carefully develop, test and implement (assuming feasible) an alternative approach to the binary nature of flag values when used in the Step 1a calculation.	Technical	High
17	3.8	Review the approaches used for dealing with • non-reportable metrics • the imputation of missing flags from individual component years (including consideration of discontinuing this practice) making the TEF documentation of these methods and approaches fully transparent.	Technical and Communications	Medium
18	3.9	Consider a formulaic approach to combining the metrics for both full-time and part-time students.	Technical	Medium
19	3.10	In the context of the different core metrics capturing a diverse range of information, consider the usefulness of a single, combined measure in Step 1a, alongside the other recommendations we make about the Step 1a process.	Wider	Medium
20	4.1.3	The Step 1a methods should be developed further, and in a holistic way, noting the other recommendations made on specific aspects, and that development should include the consideration of a net total value measure, proportion functions and approximation of confidence intervals, if possible.	Technical	High
21	4.2	Consider developing an analogous, non-benchmarked version of a combined indicator, which could be presented alongside a benchmarked version in Step 1a.	Technical	High

22	5.1	The documentation and descriptions of 'very high and very low absolute values' and their methods should be made clearer and more transparent. The appropriateness of using provider-level thresholds for each specific subject should also be reviewed.	Communications and Technical	Medium
23	5.2	Consideration should be given to removing splits with a high prevalence of small sample sizes, or at least collapsing their categories.	Technical	Medium
24	5.3	TEF users should be advised of potential small-sample-size issues when working with the subject-level data; consider making explicit reference to subjects where this is likely to be of particular concern.	Technical and Communications	Medium
25	6.1.1	To ensure harmonisation across government data, we recommend adoption of the GSS question on disability. If not, then ensure that respondents fully understand the guidelines when answering the existing question.	Technical	Medium
26	6.1.3	To improve comparability on gender identity, and to consider alongside the definition of sex: • re-label the variable to 'transgender status' to avoid confusion • ensure 'prefer not to say' is included as a standard option.	Technical	Medium
27	6.1.4	Further consideration should be given to how non-binary data on respondents' sex is treated, and the implications for data quality when binary data on sex are required for reporting purposes.	Technical	Medium
28	6.2.1	Plans and preparations should be made to handle the discontinuity caused by the forthcoming transition from SOC 2010 to SOC 2020.	Technical and Communications	Medium
29	6.2.2	On the Common Aggregation Hierarchy: • consider the name of the classification, perhaps adding 'of Academic Subjects' to its title to make it more self-explanatory, and a year to denote its introduction • plans and preparations should be made to handle the discontinuity caused by the forthcoming transition from JACS to HECoS.	Technical and Communications	Medium
30	7.1	Consult with a broad range of users on their understanding and use of existing TEF outputs, and how they would like them communicated to be as useful as possible.	Communications	Medium
31	7.2	Consider the comments given, together with user feedback, to improve the content and layout of the TEF webpages.	Communications	Medium

32	9	Consider the comments made on the appropriateness of the core metrics, and whether any improvements could be made.	Technical	Medium
33	11	Review the pilot run in TEF Year Four, and consider the usefulness of subject-level ratings, given the methods and data that support them.	Wider	High

1. Introduction

1.1 Context

Dame Shirley Pearce, who is leading the independent review of the Teaching Excellence and Student Outcomes Framework (TEF), has commissioned the Office for National Statistics (ONS), via a contract with the Department for Education (DfE), to provide an external and independent evaluation of the statistical elements of TEF. A previous review of data sources for TEF was carried out by ONS in 2016, and the present evaluation includes a brief review of progress against the recommendations made then.

Our main focus in this evaluation, however, is the statistical methods that underpin TEF and their communication, and we have evaluated these against what we consider to be statistical good practice. Useful reference material for our evaluation includes the Code of Practice for Statistics (UK Statistics Authority, 2018), although it is not the role of this evaluation to formally assess TEF against this code, and the European Statistical System's dimensions of statistical output quality: relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, and accessibility and clarity; see, for example, Quality Assurance Framework of the European Statistical System (Eurostat, no date).

This evaluation has been carried out by the Government Statistical Service's Methodology Advisory Service within ONS, and undertaken by a group of experienced methodologists, statisticians and social researchers. Our experience lies mainly in the methodology used in official statistics, and not specifically in higher education, the methods and processes of TEF, or more general approaches to measuring institutional performance. We have tried to ensure our understanding of TEF methods is correct, but any errors remain our own.

1.2 An introduction to TEF and its statistics

In its Green Paper on teaching excellence, social mobility and student choice, the then Department for Business, Innovation and Skills (BIS) proposed the development of TEF as a vehicle to "identify and incentivise the highest quality teaching to drive up standards in higher education, deliver better quality for students and employers and better value for taxpayers" (BIS, 2015). It was envisaged that TEF would change providers' behaviour, such that higher performers would be able to attract more students and raise tuition fees, while lower performers would be forced to increase standards or else face the prospect of exiting the sector. The policy aims of TEF have been outlined by BIS (2015 and 2016b), and more recently summarised by the Office for Students (OfS, 2018.44):

- to better inform students' choices about what and where to study
- to raise esteem for teaching
- to recognise and reward better teaching
- to better meet the needs of employers, business, industry and the professions.

Since its introduction in 2016, participation in TEF has been voluntary for UK higher education providers and it has been applied at provider level, which gives a single assessment and award across a whole university or college. Almost 300 universities and colleges across the UK have taken part, each receiving a Gold, Silver or Bronze award. DfE is committed to introducing the framework at a subject level. This to provide more useful information to students who tend to choose their subject of study first. The subject-level TEF is currently being piloted.

Organisational measurement is a challenging activity and is still a developing field. The work carried out to produce TEF is a contribution to this area of measurement and, as such, is to be commended.

A TEF assessment is based on:

- various quantitative data sources providing different measures thought to be associated with teaching quality, learning environment, or student outcomes and learning gain; these are called the core metrics
- other quantitative data, including breakdowns of the core metrics by various demographic sub-groups (called split metrics), as well as other data, for example, student numbers (this is called contextual information)
- other, qualitative information, for example statements written by the providers (such as universities) in their TEF submissions.

The principal, final outcome is a rating of Gold, Silver or Bronze for the provider, or subject within a provider.

There are various stages to assessment. The first (called Step 1a) is entirely formulaic, combining the information contained in the core metrics to a (usually) single starting point for the final award. Given its empirical nature, the scope of our evaluation is largely confined to Step 1a, though the split metrics used in Step 1b are also briefly considered. TEF assessors and a TEF Panel, comprising academics, student representatives and other experts, consider this (formulaic) starting point alongside the contextual data and the provider's submission in a holistic way, and use their experience and professional judgement to decide upon the final rating. Given its subjective nature, this part of the TEF process is largely beyond the scope of our statistical evaluation, though we have considered elements that are pertinent to how the statistics are used, interpreted and communicated.

1.3 About this report

1.3.1 Scope

Our evaluation focusses on the statistical properties of the data sources, metrics, techniques, and (to some extent) the outputs of TEF. The scope of the evaluation was agreed with Dame Shirley Pearce and her statistics steering group prior to its commencement (ONS 2019a) and we summarise the main points here.

The scope of our evaluation is limited to the statistical processes of TEF, with other aspects of the framework being considered by the wider independent review of TEF. Although the place of the statistical element of TEF within the context of the broader framework is also important, limitations of time have allowed us only brief consideration of this. Rather, we have concentrated on two broad aspects identified as being priorities: firstly, statistical methods and data analysis, and secondly, transparency and communication of the statistics and statistical concepts. That means that we have not covered every statistical topic, and some topics have been considered in less depth than we would have liked. Most of the evaluation has involved a consideration of the documented methodology, together with analysis of data; we have also conducted a small number of interviews with TEF assessors about their use of statistical information.

Specific topics within scope were:

- High priority:
    - statistical methods and data analysis (including – in no particular order – z-scores, normality assumptions, multiple hypothesis tests, use of flags and binary values, high and low absolute values, statistical uncertainty, missing data, timeliness aspects, benchmarking and alternatives, small samples and subgroups, combined indicators, weights, initial hypotheses, and various analyses on sensitivity and alternative approaches)
    - transparency and communication (including different user perspectives, comparability, explanations of users, uncertainty, suppression, harmonisation, and the effect of external factors).
- Medium priority:
    - updates on recommendations from the ONS 2016 review of data sources
    - review of new data sources (such as Graduate Outcomes)
    - coding and classification
    - appropriateness of TEF metrics.

We have managed to cover most of these topics to some extent in our evaluation, despite tight timescales.

Our evaluation is also largely confined to the methods currently employed by TEF, focussing our attention on what is there rather than wider questions of whether this is in some sense the correct approach (for example, we have not examined whether the metrics actually provide a good measure of what they purport to assess).

Where time has permitted, we have made suggestions for development options to check the feasibility of our recommendations, but our remit here does not include that developmental work. Our suggestions are not ready for implementation and would require further work.

We recognise that TEF has very much been designed with policy aims in mind (BIS, 2015 and 2016b), and that the methods employed within it are constrained by the requirement to support these aims. Nonetheless, we reiterate that the purpose of our evaluation is to objectively appraise the statistical processes that underpin TEF from a purely statistical standpoint. Although there may be merit in evaluating the extent to which TEF fulfils its originally stated policy aims, or even a broader cost-benefit analysis of the implementation of TEF against the counterfactual of no TEF (or an alternative type of 'TEF'), this is very much beyond the scope of our review.

1.3.2 Sources of information

For this independent evaluation, we have worked mostly from publicly available information about TEF published by DfE and OfS. This has comprised, largely, three documents:

- DfE (2017a): Teaching Excellence and Student Outcomes Framework Specification, October 2017
- OfS (2018.44): Teaching Excellence and Student Outcomes Framework: Subject-level pilot guide, 22 October 2018
- OfS (2018.44a): Teaching Excellence and Student Outcomes Framework: Guide to subject-level pilot data, 22 October 2018.

A number of other documents from DfE and OfS have been consulted too, as well as other sources (such as the Unistats and Wonkhe websites). We have also considered the Royal Statistical Society's statements on TEF (RSS, 2018 and 2019), the latter of which comments on nine main areas of the TEF process, including uncertainty handling, comparability, benchmarking and transparency, and small sample sizes.

We have checked our understanding of topics with DfE and OfS where something has been unclear. In places, our descriptions omit the finer details of methods where these do not add value to the comments we make. As such, this report should not be considered as providing a complete or definitive description of TEF statistical methods.

1.3.3 Analytical work

We have used publicly available data for much of our analysis, and have also been provided with anonymised data from the subject-level pilot by OfS. The following data were provided to us:

- TEF Year Three: publicly available metrics and final awards data, with OfS data used to reconcile our derived Step 1a starting points with those that were actually calculated by OfS (2017 method)
- TEF Year Four: publicly available metrics data, but methods as per TEF Year Three (2017 method)
- subject- and provider-level pilot (2018 method): data from OfS.

Note that our analysis of the published TEF Year Three data has focussed on the 86 providers that were ultimately assessed and assigned a final TEF award. This is in contrast with the published TEF Year Four data, for which metrics are available for a much broader group of providers (DfE has confirmed this to be all providers in England within the scope of TEF, irrespective of whether they were assessed for a TEF award, and those from the devolved administrations that chose to participate; full details can be found in the notes sections of the TEF workbooks). This discrepancy is reflected in our analysis.

1.3.4 Qualitative analysis

We have carried out a small number of interviews with relevant parties, notably the interviews with four TEF assessors and Panel members

1.4 Structure of this report

In Section 2 we consider the core metrics and the methods and assumptions required for their creation, outline the data and methods used to create the Step 1a starting point, and do this for both the main methods currently in use and those being piloted. We provide some analysis of the distribution of these metrics across providers.

In Section 3 we consider how the core metrics are combined in the Step 1a process, particularly around the use of the TEF indicators, benchmarks, z-scores, flags, and weights. We identify a number of limitations with the current statistical process and make several recommendations on how it could be improved, some of which will require further development, research and consideration.

In Section 4 we proceed to test some of our suggestions on real TEF data, and consider what the results suggest about the robustness of the process.

In Section 5 we consider various aspects of the contextual data, alongside some more general statistical aspects of the data and processes.

In Section 6 we make wider statistical comments on harmonisation and classification.

In Section 7 we look at communication, specifically clarity and transparency.

In Section 8 we summarise developments since the ONS 2016 review.

In Section 9 we briefly consider the appropriateness of the TEF metrics in the context of the National Audit Office's 'Choosing the Right FABRIC' document (NAO, 2001).

The scope of our report largely ends at that point, as that is the final point where formulaic approaches are used, and the point at which academic, professional judgement starts to be applied. However, we have interviewed four assessors about their use of the statistical information, and report on our findings in Section 10. We have also considered what other sources of information might be useful for inclusion either as contextual information or in the core metrics.

In Section 11 we provide a summary of our conclusions.

Finally, we note that TEF and its methods utilise many technical terms; where possible, we have adopted the same terminology in this report as is used the main DfE and OfS documentation. We attempt to explain those as-and-when they are introduced in the report, but also provide a glossary of the commonly used terms along with references at the end of this report.

1.5 Acknowledgements

The ONS team would like to thank colleagues from DfE and OfS for their support throughout this work; they have provided helpful additional documents and answered many questions. Thanks also to members of Dame Shirley Pearce's advisory group who provided feedback on an early version of this report. We were also helped by staff from the Higher Education Statistics Agency and many other colleagues at ONS – our thanks go to all of them.

1.6 About the authors

Gareth James: Gareth has a Methods Lead role in the Methodology Division of the Office for National Statistics, and currently heads the Research and Policy Analysis hubs within the division. He has worked at ONS since 2001 and has led Sample Design & Estimation and Statistical Computing branches, as well as spending some time working in the National Accounts area. Gareth has a BSc in Mathematics and a PhD in Statistics from Cardiff University.

Daniel Ayoubkhani: Daniel is a Senior Methodologist at the Office for National Statistics, where he has worked for nine years in both methodological and production roles. Daniel also has around four years' experience in the private sector, as a Senior Modelling Analyst at a major financial services provider and as the Principal Statistician in a commercial healthcare research consultancy. Daniel has a BSc in Economics from Cardiff University and an MSc in Official Statistics from the University of Southampton, achieving the latter with distinction and two Dean's List awards.

Jeff Ralph: Jeff joined ONS in 2004, initially managing statistical modernisation projects before moving to Methodology to work on Index Numbers and Price Statistics. From 2014 to 2018 he managed the Methodology Advisory Service and led the review of TEF data sources in 2016. Jeff was a member of the Advisory Panel on Consumer Prices from 2016 to 2018 and was appointed Royal Statistical Society William Guy Schools' Lecturer for the academic year 2017-18. He has co-authored two books: "A Practical Introduction to Index Numbers" and "Inflation: History and Measurement". Jeff retired from the Civil Service in March 2018 and now has visiting academic status at the University of Southampton. Jeff has a BSc in Mathematics and Physics, an MSc in Applied Maths and a PhD in Theoretical Physics from the University of Warwick.

Gentiana D. Roarson: Gentiana is a Senior Researcher in Best Practice and Impact Division of the Office for National Statistics. She is coordinating the Methodology Advisory Service and Government Statistical Service Methodology Advisory Committee (GSS MAC) redesign to better meet the needs of the GSS. Previously, she led the National Statistician's Quality Review of Privacy and Data Confidentiality Methods. She has an MSc in Forensic Psychology and has previously worked in Data Collection Methodology, ONS and Knowledge and Analysis Services, Welsh Government.

2. Overview of the methods used for core metrics

This section provides an overview of how the core metrics are used in the creation of the Step 1a starting point. It includes brief descriptions of the variables or measures associated with each metric, the metrics themselves, their weights, flags and the starting-point diagrams under the two methods currently in use or being piloted.

2.1 Measures associated with the core metrics

We first introduce the set of measures associated with the Teaching Excellence and Student Outcomes Framework (TEF) core metrics. We do not dwell on the metrics themselves (that is, on what they measure), but focus instead on their statistical properties. Descriptions of what they measure have been documented by the Department for Education (DfE, 2017a, page 32) and the Office for Students (OfS, 2018.44a, page 13). Between them, the core metrics contribute to measures considered to be associated with three aspects of quality – teaching quality, learning environment, and student outcome and learning gain – and include topics such as satisfaction with teaching and entering employment or further study after graduation.

Each core metric from each provider or subject-within-provider has a number of measures associated with it, as can be found in, for example, the publicly available TEF workbooks. Table 1 provides a summary of these measures, and we then describe each measure in turn.

Table 1. Measures associated with each core metric of a provider or subject

Name of measure	Symbolic notation	Type of variable	Unit of measurement	Bounds or categories	
indicator	p	continuous	percentage	[0, 100]	
benchmark	E	continuous	percentage	[0, 100]	
difference	d	continuous	percentage points	[-100, 100]	
standard deviation of difference	std(d)	continuous	percentage points	[0, infinity)	
z-score	z	continuous	none	(-infinity, infinity)	
flag		categorical	not applicable	--	usually grouped to negative
				-	
				=, neutral, none or unflagged	
				+	usually grouped to positive
				++	
				not reportable (for various reasons)	

Indicator (p): this is a percentage in every case, and represents something positive or desirable. Examples include the percentage of students who reported being satisfied with the teaching on their course, or those who went on to achieve employment or further study. The indicator may be based on a survey of students, such as the National Student Survey (NSS) or the Destination of Leavers of Higher Education (DLHE) survey, or from administrative data.

On the NSS (at least), a five-point Likert or rating scale is used, with answer categories Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, and Strongly Agree (see, for example, NSS 2017 Core Questionnaire). The latter two categories are combined to derive the 'percentage positive' score that forms the indicator. We note that the NSS scale employs a middle category, about which the literature seems to have mixed views.

Benchmark (*E*): this calculated as the average indicator across all providers applied within each benchmark group that is then aggregated back to the provider in question taking account of that provider's student- and subject-mix. Benchmarking is intended to allow a more meaningful interpretation of the indicator, and fairer comparisons to be made across providers. For each metric, the benchmark represents each provider's expected indicator value had it performed as the sector average, given its student and subject mix. Note that the benchmark is specific to the provider; different providers have different benchmark values for the same metric.

Difference (*d*): this is defined as the value of the indicator minus the value of the benchmark and is measured in percentage points. In broad terms, a positive difference can be regarded as a more-positive-than-expected outcome for a provider in this metric, and a negative difference as less-positive than expected. The question then is whether this observed difference has occurred because of some artefact of the provider, or whether it might have occurred by chance relating to the students who happened to be at that provider at the time in question; the magnitude of the difference is obviously important too.

z-score (*z*): this is used to assess how likely the observed difference is to have occurred by chance, and uses the standard deviation (or, more correctly, the estimated standard error) of the difference as an input to its calculation. Further attention is paid to this in Section 3.4.6.

Flag: depending on the value of the difference and the z-score (and occasionally the benchmark itself if especially high), whether or not a provider's metric is flagged is determined by a set of rules. A realised flag can be interpreted as showing statistically significant and notably large positive and negative differences from the benchmark:

- single-positive flag ('+'): $d > 2$ and $z > 1.96$, or just $z > 1.96$ if $E > 97$
- double-positive flag ('++'): $d > 3$ and $z > 3$, or just $z > 3$ if $E > 97$
- single-negative flag ('+'): $d < -2$ and $z < -1.96$
- double-negative flag ('++'): $d < -3$ and $z < -3$

Note that unflagged metrics are sometimes shown as '=', and other symbols denote "not-reportable" and other statuses.

Broadly, flags have two purposes:

- the visual identification of provider and subject metrics where the difference between the indicator and the benchmark is materially and significantly different from zero
- as part of the Step 1a process of generating starting points; each flag carries a specified value.

We also define *w* as the **weight** given to a particular metric. The weights are not shown in the TEF workbooks, but are quoted in the TEF documentation (for example, DfE (2017a, page 57) and OfS (2018.44a, page 14)). The values of *w* are metric-specific, and do not depend on the provider or subject. The weights are used when combining metrics into a single outcome. Section 2.2.2 contains more detailed considerations of the weights.

2.2 Current approaches to combining core metrics to form the starting point in Step 1a

2.2.1 The 2017 and 2018 methods

Before describing how the core metrics are combined, we note there are two sets of metrics and associated methods currently in use. We define those as the 2017 and 2018 methods:

> **The 2017 method:** the first method is described in DfE (2017a); this covers the provider-level TEF process used in Year Three (2017/18 academic year), and continues to apply for Year Four (2018/19 academic year) provider-level data except for various updates, which are mainly outside the scope of this statistical evaluation.
>
> For brevity, we refer to this as the 2017 method, but reiterate that it applies to provider-level assessment in both the 2017/18 and 2018/19 academic years (Years Three and Four of TEF, respectively).
>
> **The 2018 method:** the second method, which applies to the subject-level pilot (also called the second pilot, Pilot 2, or pilot in the 2018/19 academic year) refers to the methods used for the subject-level assessments and how these assessments combine to form the provider-level assessment. These methods are documented in OfS (2018.44 and 2018.44a), and are largely similar-in-principle to the provider-level methods, but differ in terms of the core metrics that are included, and in their processing.
>
> For brevity, we refer to this approach as the 2018 method, but reiterate that it applies only to the subject-level pilot and not the usual, Year Four TEF assessments.

We regard the 2018 method as being a development and refinement of the 2017 method.

2.2.2 The core metrics and their weights

The 2017 method uses six core metrics, and the 2018 method uses nine – these are named in Tables 2a and 2b. Five of the core metrics are common and feature in both the 2017 and 2018 methods. However, we note that the detail of the composition of those common metrics, in terms of the individual survey questions that might comprise them, has not always remained the same. OfS (2018.44, page 81), for example, shows differences in the composition or question wording of the NSS metrics between TEF in 2016 and TEF in 2017 and 2018.

Weights are attributed to each core metric, and have been revised for the 2018 method, a necessity in any case to accommodate the increase in the number of core metrics from six to nine. Tables 2a and 2b summarise the core metrics and their absolute weights (a mix of 0.5 and 1.0 per metric, which sum to 4.5 in the 2017 method, and a mix of 0.5, 1.0 and 2.0 per metric, which sum to 7.5 in the 2018 method); we have converted the absolute weights to relative weights (percentages, which sum to 100% across all metrics) to allow for easier comparison.

Table 2a. Metrics and their weights used in the 2017 method

Teaching quality	Learning environment	Student outcomes and learning gain
The teaching on my course (NSS): 0.5 = 11.1% Assessment and feedback (NSS): 0.5 = 11.1%	Academic support (NSS): 0.5 = 11.1% Continuation (HESA): 1.0 = 22.2%	Employment / further study (DLHE)*: 1.0 = 22.2% Highly skilled employment / further study (DLHE)*: 1.0 = 22.2%
1.0 = 22.2%	**1.5 = 33.3%**	**2.0 = 44.4%**
Total NSS weight is 1.5 = 33.3% **Total weight is 4.5 = 100%**		

** Metric not present in the 2018 method*

Table 2b. Metrics and their weights used in the 2018 method

Teaching quality	Learning environment	Student outcomes and learning gain
The teaching on my course (NSS): 0.5 = 6.7% Assessment and feedback (NSS): 0.5 = 6.7% Student voice (NSS)*: 0.5 = 6.7%	Academic support (NSS): 0.5 = 6.7% Learning resources (NSS)*: 0.5 = 6.7% Continuation (HESA): 2.0 = 26.7%	Highly skilled employment or higher study (DLHE): 1.0 = 13.3% Sustained employment or further study (LEO)*: 1.0 = 13.3% Above median earnings threshold or in higher study (LEO)*: 1.0 = 13.3%
1.5 = 20.0%	**3.0 = 40.0%**	**3.0 / 7.5 = 40.0%**
Total NSS weight is 2.5 = 33.3% **Total weight is 7.5 = 100%**		

** Metric added in the 2018 method*

2.2.3 Starting points and Step 1a

It is only core metrics that are flagged (those given singly or doubly positive or negative flags) that contribute any values to the starting-point calculation in Step 1a of the process. The weights of metrics with positive flags are summed, and, likewise, the weights of metrics with negative flags are summed. A diagram – an easy-to-visualise interpretation of a set of somewhat complicated rules – is then used to determine how that combination of total positive values and total negative values translates to a Step 1a starting point. We note that the 2017 method has only Gold (G), Silver (S) and Bronze (B) as possible starting-point categories, but that the 2018 method introduces additional borderline categories of Gold/Silver (G/S) and Silver/Bronze (S/B).

Figure 2 reproduces the diagram for the 2018 method found in OfS (2018.44, page 63). The 2017 method doesn't have a similar diagram in DfE (2017a), so the one illustrated in Figure 1 has been derived using the published ruleset. To allow easy, visual comparison between the two, we have added the relative weights from Tables 2a and 2b to the positive and negative axes, and present the two diagrams on approximately the same scale (the widths of the two diagrams, representing 100% of the available positive value, are approximately the same, and likewise the heights for negative values).

Figure 1. 2017 method starting-point diagram

Starting points for Gold (G), Silver (S) and Bronze (B)

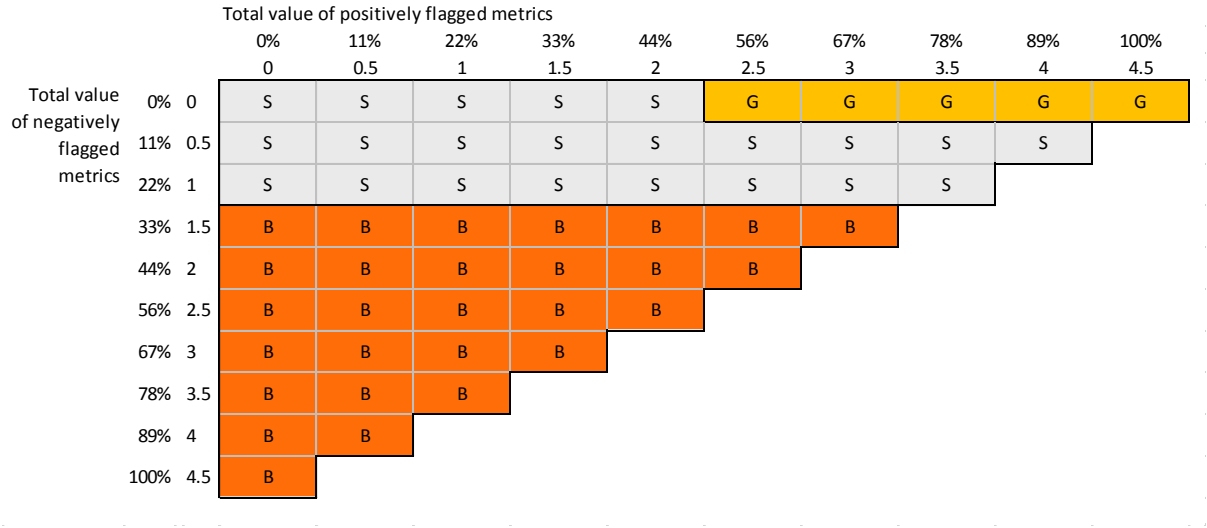
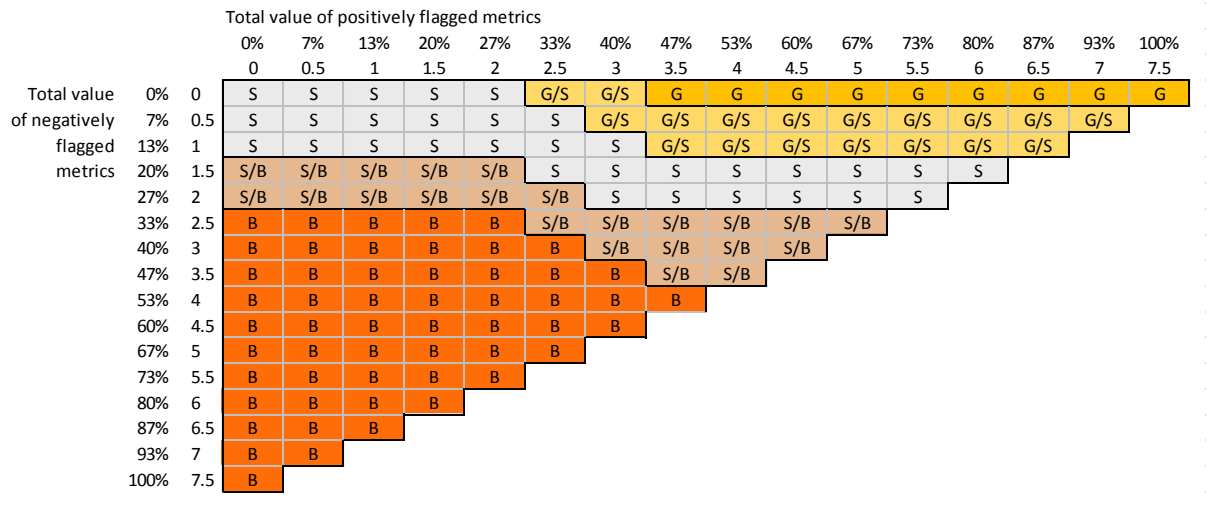


Figure 2. 2018 method starting-point diagram

Starting points for Gold (G), Silver (S) and Bronze (B), and the borderline ratings of Gold/Silver (G/S) and Silver/Bronze (S/B)



A very simplistic comparison of 2018 with 2017 reveals that Gold now stretches a little further left (at 47% of the available positive value, compared with 56%, and into comparable, previously Silver territory), and the new Gold/Silver borderline category surrounding it also covers previously Silver territory. Given that a Gold starting point can only be obtained by having no negatively flagged metrics, one could argue that is more difficult to avoid negatives with nine metrics than with six, but correlations between metrics makes this evaluation more difficult to quantify.

Similar observation reveals Bronze starts at about the same depth (33% of the available negative value), but the upper threshold decreases further to the right, potentially pushing previously Bronze providers more towards Silver. The Silver/Bronze borderline category sits above it, in previously Silver territory. As such, the Silver territory has been shrunk both above and below.

We reiterate that these comparisons are very simplistic and rely on various assumptions. Our intention at this stage is to objectively describe and compare the 2017 and 2018 approaches, without casting aspersions over their relative merits or otherwise. Further, we stress that these comments relate only to the Step 1a, formulaic starting points, and that the TEF assessors and Panel reviewers can change these starting points in light of other data and expert opinion.

3. Evaluation of current Step 1a methods and data

Having provided an overview and brief description of the Step 1a process in Section 2, we now move on to a more critical appraisal of the methods, data, and various aspects of the process.

Please note that whenever methodology relating to 'providers' is discussed in this section, it should be taken as also applying to 'subjects-within-provider' (as under the 2018 method). All empirical analysis presented in this section relates to provider-level data for the Teaching Excellence and Student Outcomes Framework (TEF) Years Three and Four using the 2017 method; analysis of the subject-level 2018/19 pilot TEF data can be found in Section 5.3.

The TEF metrics are based on a variety of data sources; some are from survey data, whereas others are derived from administrative data. We do not dwell on the quality of the data sources themselves in this evaluation, but note that quality is an important consideration, and is something that the 2016 TEF Review of Sources (ONS, 2016) considered. The Code of Practice for Statistics (OSR, 2018) notes that quality should be monitored and reported regularly and the extent and nature of uncertainty be clearly explained (Q3.3), and recommends that systematic and periodic reviews of data should be undertaken (Q3.5).

3.1 Reference period: timeliness, comparability and stability

The core metrics (and other data) presented in TEF outputs and used in the assessments are based on the most recent data available. As noted, the various core metrics of TEF are derived from a variety of data sources, and the timeliness of these sources varies. Therefore, the reference period – that to which the data relate – varies from metric-to-metric, and likewise so do the population(s) of students and former students to whom the data pertain. The 'Continuation' metric, for example, relates to students who continue their studies from their first year into their second, the National Student Survey (NSS) collects information about final-year students, whereas the Destination of Leavers of Higher Education (DLHE) survey and the Graduate Outcomes (GO) survey refer to students after graduation, and some of the information about employed graduates takes longer still to arrive via the tax system. In the following sections of this report, we make further comments about the intended reference population of TEF.

The core metrics for most providers are based on the 'most recent three years' of available data; that is, the summary measures are calculated across all students or graduates using the most recent three years of available data at the time of calculation. Analysis conducted by the Office for Students (OfS, internal report) suggests that calculating the indicators over a three-year window is necessary to smooth out the volatility that is inherent in single-year estimates; there is, therefore, a trade-off between accuracy and other measures of quality (various aspects of timeliness, coherence, comparability and relevance). That said, not all providers have a full three years of historical data available, in which case data from shorter time periods are used instead.

In our evaluation, we have not investigated the stability of TEF assessments over time. At present, that would be difficult to do in a meaningful way, since:

- TEF has not been in existence for very long, thus only very short time series are available
- TEF methods have changed, for example benchmarking factors were altered between TEF Years Two and Three
- those providers assessed in TEF so far have (usually) been assessed only once
- those providers assessed in any given year (for example, TEF Year Three) were not chosen at random, and would likely be considered unrepresentative of providers more generally.

However, it should be possible to consider the stability of the input metrics instead, although that, along with other aspects relating to the data sources, is outside the scope of this evaluation. Longer, stand-alone time series of the component metric indicators (for example, from the National Student Survey (NSS)) could be considered and analysed for each provider: these series should be available for each year, even if a TEF assessment and rating of the provider is not undertaken every year. It may be useful for longer time series and analyses to be made available where possible.

Recommendation 1: Time series analysis, including an assessment of stability, of TEF-input core-metric indicator series should be conducted and made available on an ongoing basis.

3.2 The indicators

All TEF metrics' indicators are presented as percentages, and as something seen as positive or desirable (rates of continuation are used, for example, rather than rates of non-continuation).

3.2.1 Statistical uncertainty in the indicators and super-populations

The indicators are just point estimators. That is, they provide a single value considered in some way to represent the 'best estimate' or 'most likely' value of some unknown or underlying true parameter.

However, for both survey and non-survey data, there exists statistical uncertainty, meaning that the estimates may not be accurate or precise measures of the quantities that they aim to represent. Although the NSS and DLHE survey are censuses, they suffer from non-response; administrative data may suffer from incompleteness. All sources of data are, inevitably, subject to various measurement-error effects too, which can result in inaccurate information being recorded.

Estimates from the NSS, for example, have confidence intervals available (these are contained in spreadsheets available on the OfS website), which should help convey some of that uncertainty. Examination of those confidence intervals reveals them to be asymmetric (reflecting the nature of estimated proportions not near 0.5). A note on the OfS website suggests these could, among other options, be Wilson score intervals, and OfS has confirmed to us that Wilson score intervals are indeed calculated. OfS also provided a document (subsequently found available as Appendix 2 in UK Centre for Materials Education) that quotes a precise formula for the confidence interval, stating it has been modified to contain an 'adjustment for the false discovery rate for multiple comparisons'. That adjustment seems to be the use of 2.17 as the z-value (as opposed to 1.96, for example), but precisely how that number was derived is not stated. We could not exactly recover the NSS confidence intervals quoted in the available spreadsheet by use of that formula, even after allowing for rounding, but didn't find any confidence limits that differed by more than one percentage point.

The confidence interval calculations seem to be based upon the proportion of students responding positively (that is, answering as 'mostly agreeing' or 'definitely agreeing' (categories 4 and 5) with the question's statement) defined as *p*, and the column headed 'response' (presumably the number of responding students) used as *n*. The latter suggests an assumption of some super-population, rather than the current student population cohort, as the target population. Thus, statistical uncertainty would still exist even if a fully-responding census were realised (as is sometimes the case for particular groups), and it corresponds with the notion of the actual students in a provider's population being just one (random) realisation of many other populations of students who could have attended that provider, or may do so in the future.

The super-population approach seems reasonable, and it is something that is applied in other situations. However, it differs from many other official statistics (such as those usually produced by the Office for National Statistics (ONS)), in which a fixed and finite population is assumed, so that a fully responding census would result in the point estimator having a confidence interval of zero width. That approach also has merits: it gives an indication of the possible error incurred by selecting (or realising) only a sample of the target population rather than all of it. If one were interested only in the actual population of students present at a particular provider at a particular time, then the finite-population approach would be appropriate.

As discussed in more detail later in Section 3.4.2, the super-population is also the target population of TEF (though this does not seem to be explicitly stated in the publicly available documentation). There is therefore some consistency between the assumptions in both the aggregate TEF metric methods and in their underlying data sources with respect to the calculation of their confidence intervals.

Recommendation 2: Documentation describing the NSS confidence-interval calculations, and the assumptions upon which they are based, should be updated and made clearer.

Although confidence intervals have been calculated, for some metrics at least, we note that the statistical uncertainty embodied by those intervals does not feature directly in the TEF calculations, although sampling variation is captured in the aggregate TEF metrics through the z-score.

That the confidence intervals associated with the metrics' indicators are also absent from sight in the TEF workbooks, might give the impression that the indicators are more accurate than is really the case. We also understand that the NSS indicators are unweighted proportions of the responses received. It would usually be good practice to weight survey data to help reduce potential non-response bias, and this was recommended in ONS (2016). OfS has analysed response patterns in the NSS and tested the effect of introducing non-response weighting (HEFCE, 2016); the conclusion was that non-response weighting would have little effect at this time on NSS indicators. It would be good practice (UK Statistics Authority, 2018) to continue checking this regularly. Similar practice should be employed on other data sources too.

Recommendation 3: Monitor and report on potential non-response bias in the NSS (and other data sources) on a regular basis, introducing appropriate non-response weighting or calibration if required.

Without doubt, statistical uncertainty exists in the TEF metrics data. It is not fully taken into account in Step 1a methods, nor is any indication of it displayed in the provider-level TEF workbooks. We consider statistical uncertainty, and its possible impacts, further throughout Section 3 and later in Section 7. We make various recommendations later about improving communication of statistical uncertainty in TEF; these should aim to improve transparency and reduce the risk of users of TEF data mistakenly assuming greater accuracy than is warranted.

3.3 Benchmarking

3.3.1 General approach

The need to control for input mix when assessing institutional outputs was raised by Goldstein and Spiegelhalter (1996), particularly with respect to constructing 'league tables' for comparing performance. The same philosophical approach is employed in TEF (although it is implemented in a design-based manner rather than a model-based one, as proposed by Goldstein and Spiegelhalter (1996)). One point of departure is that TEF is not intended to be a ranking tool; however, benchmarking introduces a dependency between each individual provider's performance and that of the other providers in the sector such that the resulting flags are based on relative performance. This says nothing about absolute performance, for example a provider might be performing below its UK sector average for a particular metric, but what if the UK higher education system is particularly strong in this dimension of quality when compared internationally, or if there is a secular trend of improvement among all UK providers? The individual provider will not be rewarded in either of these situations, because their performance is assessed against a benchmark that is determined by the performance of other UK providers in their sector.

The current TEF approach to benchmarking poses the question 'What would the provider's indicator have been had it performed as per the sector average, given its student and subject mix?' We note that this is just one possible approach and alternative formulations also exist, for example 'What would the provider's indicator have been had it had the same student and subject mix as the sector average, given its performance?' There is no single 'correct' approach, and different approaches will often lead to different values of the benchmark, E, and therefore also the difference from it, d. It is unclear whether any sensitivity analysis has been performed to investigate the impact on the Step 1a outputs of changing the overall approach to benchmarking. However, we also note the equivalence of some of the available methods. For example, TEF adopts a design-based approach; alternatively, a model-based approach – where the benchmark factors and all of their higher-order interactions are included as fixed effects, and the student-level predicted values are then averaged to provider-level – could have been used, but these two formulations are equivalent.

We note that Recommendation 5 of the ONS 2016 review was to "carry out a methodological review of the creation and use of benchmarks" (ONS, 2016). At the time of writing this report, OfS is in the process of reviewing the existing benchmarking methodology (independently of our evaluation), including assessing its sensitivity and calibration to certain assumptions, and examining a range of options for extending it. OfS also commissioned an external agency, Alma Economics, to explore alternative benchmarking methodologies, the report for which has since been published (OfS, 2019a). Further information on progress against the recommendations of the ONS 2016 review can be found in Section 8 of this evaluation.

3.3.2 Selection of the benchmarking factors

In this section we evaluate the mechanics and implications of the statistical process by which the benchmarking factors are selected, and identify some potential improvements to this process. This includes consideration of variable selection techniques, the possibility of sparse benchmarking groups and the issue of self-benchmarking (which is analysed empirically), and the metrics that may be reported relating to the goodness-of-fit of the benchmarking models.

In general terms, the benchmarking factors are selected according to a set of guiding principles (OfS, 2019b) that relate to the factors' coverage and scope, their statistical properties, and their data quality and availability; for example, benchmarking factors should be outside the control of providers. From a purely statistical viewpoint, these appear reasonable and clearly some thought has gone into their development, and we welcome the fact that these principles are pre-specified and in the public domain. More specifically, the approach to selecting the benchmarking factors involves fitting statistical models to student-level data for each of the metrics, split by full- and part-time provision. Through discussion with colleagues at OfS, the model-building approach seems to be a fairly standard, but thorough, one that involves a mixture of empirical rules-based criteria and more subjective concerns to selecting the relevant covariates; 'as much an art as it is a science', as is often said of such practical exercises. Formal selection criteria include performing backward deletion on covariates with p-values less than 0.05; although common practice, the pitfalls of this approach are well-documented (for example, inflation of the parameter estimates in the final model). The large samples on which the models are built will tend the p-values to 0, even though some of the variables in the model may be adding very little explanatory power. OfS might consider refining its statistical approach to variable selection and investigate shrinkage techniques such as the Least Absolute Shrinkage and Selection Operator (LASSO). Large sample sizes would also permit some form of cross-validation, which would provide some protection against overfitting and allow OfS to more formally evaluate the bias (flexibility) versus variance (generalisability) trade-off inherent in their models; such a practice is not currently undertaken.

The cross-tabulated grid of benchmarking factors and providers may be relatively sparse for certain metrics (or at least concentrated more in certain places of the grid than others). For example, in TEF Year Three, more than half of the cells were empty for the 'Highly skilled employment or further study' metric (57% for the full-time and 74% for part-time). Furthermore, the effect of small sample sizes within cells raises questions over the precision of the estimated benchmark values *E*; in TEF Year Three, the proportion of populated cells containing fewer than 15 students ranged from 42% ('Employment or further study') to 72% ('Highly skilled employment or further study') for full-time provision, and 29% (all three NSS metrics) to 82% ('Highly skilled employment or further study') for part-time provision.

The multi-dimensional nature of the benchmarking grid also raises the possibility of having a small number of providers in a cell and subsequent self-benchmarking (whereby a provider makes a large contribution to its own benchmark). This appears to not be a problem for most provider-metric pairs (see Tables 3a and 3b), where the mean contribution ranged from 3.3% ('Academic support' and 'Assessment and feedback') to 4.9% ('Highly skilled employment or further study') in TEF Year Three; and from 2.7% ('Employment or further study') to 5.8% ('Highly skilled employment or further study') in TEF Year Four.

However, there were some instances where individual providers made up a relatively large percentage of their own benchmark: in TEF Year Three, 21 provider-metrics contributed more than 10%, three contributed more than 20%, and the largest contribution was 34.5%; and in TEF Year

Four, 145 provider-metrics contributed more than 10%, 39 contributed more than 20%, and the largest contribution was 78.3%. Cases such as these make it more difficult for the provider to be flagged positively or negatively against a particular metric (as shown in Forster (no date)), and result in the provider tending towards Silver at Step 1a.

Table 3a. Summary statistics on providers' contribution to their own benchmark* (%), core metrics, TEF Year Three

Metric	Min	Q1	Median	Mean	Q3	Max
Academic support	0.3	1.5	2.6	3.3	4.0	14.7
Assessment and feedback	0.3	1.5	2.6	3.3	4.0	14.7
Continuation	0.1	1.7	2.7	3.9	4.2	34.5
Employment or further study	0.5	0.9	1.8	2.5	2.7	17.2
Highly skilled employment or further study	0.4	2.2	3.9	4.9	6.1	24.5
The teaching on my course	0.3	1.6	2.6	3.4	4.2	14.7

** As quoted in published TEF statistics. This is a weighted average of the provider's student size to the overall (total) student size in the domain of interest (for example, subject)*

Table 3b. Summary statistics on providers' contribution to their own benchmark* (%), core metrics, TEF Year Four

Metric	Min	Q1	Median	Mean	Q3	Max
Academic support	0.0	1.5	2.4	4.0	4.6	78.3
Assessment and feedback	0.0	1.4	2.4	3.9	4.6	78.3
Continuation	0.1	1.8	2.7	4.1	4.2	72.5
Employment or further study	0.1	1.0	1.5	2.7	2.9	51.7
Highly skilled employment or further study	0.1	2.4	3.9	5.8	6.4	63.3
The teaching on my course	0.0	1.4	2.4	4.0	4.6	78.3

** As quoted in published TEF statistics. This is a weighted average of the provider's student size to the overall (total) student size in the domain of interest (for example, subject)*

Studentisation (removing each provider from the calculation of its own benchmark) is a possible solution to the self-benchmarking problem, and is stated as such in Draper and Gittoes (2004), but has not been implemented in TEF. We recommend that further consideration be given to this approach, including its implications for other parts of the process (for example, any modifications needed to the derivation of the standard deviation of *d*) and the possible impact on robustness because of small sample sizes (for example, where benchmarking groups already contain a small number of providers).

Recommendation 4: Further consider Studentisation (the removal of the contribution of each provider from its own benchmark) in the benchmarking process, including its implications for other parts of the process and the possible impact on robustness due to small sample sizes.

The issue of self-benchmarking may also be remedied by reducing the size of the grid, for example by employing some of the variable reduction techniques mentioned earlier (LASSO, cross-validation), and perhaps further collapsing the bins for some of the benchmarking factors. For example, 'subject of study' has 22 bins associated with it even after collapsing to Level 1 of the Common Aggregation Hierarchy, or 33 before; it is unlikely that such granular binning is adding much to the discriminatory power of the models.

Entry qualifications is a candidate benchmarking factor in all of the models, but we note that it was not included in the final models for the NSS metrics in TEF Year Three. Presumably this decision was taken on empirical grounds, but intuitively it would seem that entry qualifications could be a reasonable proxy for student expectations, and therefore an important factor to control for, especially in the case of student-reported experience.

The models used to select the benchmark factors are logistic mixed-effects models, with fixed effects for the benchmarking factors and random intercepts at the provider level. More could be made of model estimates and diagnostics such as intraclass correlation (the proportion of total variance explained at the provider-level) and pseudo-R^2, which would shed some light on the relative size of *d* after the benchmarking factors have been controlled for; OfS should consider publishing such measures in their user guidance. We also note that the resulting provider-level residuals are themselves an estimate of *d*, so the models could in principle be used directly without the need to translate their inferences into a multi-way grid of benchmarking factors and providers, as is done in the current design-based approach. This unified approach to estimation of *d* (and its standard deviation and the resulting z-score) has appeal, but may be less transparent to lay users (though it would be impossible for interested parties to recreate the existing approach without access the student-level data in any case).

In general, although the way in which the benchmark values (*E*) are calculated is well-documented, documentation on the process for selecting the benchmark factors does not appear to be in the public domain. We therefore recommend that an overview of the process is included in the published TEF user guidance.

Recommendation 5: Consider refining the benchmark model building process, for example by investigating variable reduction techniques and cross-validation; and summarise this process in the TEF user guidance, including appropriate model outputs and diagnostics indicative of goodness-of-fit (such as pseudo-R^2).

3.3.3 Causality and attribution

We now consider the rather more general issue of how each provider's difference, *d*, should and should not be interpreted, which stems from the possibility that the benchmarking factors do not include all variables that should ideally be controlled for.

Randomised control trials are often considered to be the most reliable way of obtaining causal inferences. Clearly this approach (that is, assigning some students of a provider to a 'treatment' group which receives the provider's usual standard of education, and others to a 'control' group which receives education to a common prescribed standard) would be unethical and impractical to implement in this situation. The quantitative part of the TEF process (Step 1a, and in part Step 1b) is therefore based on observational data alone. Hence the 'value added' by the education process within each provider is inferred as the residual (*d*) between observed inputs (*E*) and observed outputs (*p*) rather than being measured directly. We can think of the residual between education inputs and outputs as comprising of three components:

- unobserved or unmeasurable differences in teaching quality, learning environment, and student outcomes / learning gain between each provider and the sector average (that is, what TEF is aiming to capture)
- remaining differences in inputs between each provider and the sector average (that is, confounding factors that meet the TEF benchmarking principles but have not been included as benchmarking factors because they are unmeasurable or (at least as of yet) unobserved
- an irreducible random error (that is, noise in the system).

TEF attempts to estimate a counterfactual outcome (what would each provider's metrics have looked like if they had performed as per their sector average, given their inputs), but the presence of (ii) in addition to (i) means that we cannot attribute the difference between each provider's indicator and its benchmark solely to process quality. In other words, the TEF metrics do not permit causal inference. It is impossible to quantify the extent to which this is a problem, purely because the three components listed above are unobservable (if they were then they could be accounted for). The published TEF documentation (DfE, 2017a) does state that the difference between *p* and *E* is comprised of differences in process quality and other factors, but this statement should be made more prominent, and the implications should be spelt out more clearly for lay users to whom this issue may not be immediately obvious.

Recommendation 6: In the published TEF documentation, make it clearer that the TEF outputs cannot solely be attributed to teaching quality, learning environment, and student outcomes / learning gain, because the benchmarking process does not take into account all confounding variables (those that would meet the TEF benchmarking principles but have not been included as benchmarking factors because they are unobserved or unmeasurable).

We also acknowledge that TEF is *informed* by statistical metrics rather than being *driven* by them; the subsequent holistic judgement of the TEF Panel is meant to place the metrics into a broader context alongside other quantitative and qualitative indicators of process quality. However, it is not clear as to the extent to which the Panel members themselves consider the non-causal nature of the metrics in their decision-making process. Furthermore, the qualitative information used by the Panel largely takes the form of provider submissions, which again involves (non-statistical) inference in the face of potentially confounding factors. We note that at no point are phenomena such as teaching quality directly observed or assessed (such as by trained experts visiting providers and observing the teaching and teaching materials) during the TEF process, as is done in the case of school and prison inspections.

Estimation of *d* relies on having all potential confounders (or at least those that meet the TEF benchmarking criteria) captured as benchmarking factors. This will not be possible because of unmeasurable factors, such as those that are idiosyncratic to individual students, but considering a wide range of observable datasets will be beneficial. At present, the benchmarking process makes use of a collection of linked datasets (for example, the NSS, Higher Education Statistics Agency (HESA), Individualised Learner Records), and we commend OfS for exploring new data sources such as the National Pupil Database and student loans data. We encourage continued investigation of new data sources from which candidate benchmarking factors may be obtained, and look forward to seeing the results of OfS's current investigations. However, a number of important factors that almost certainly explain some variations in the TEF metrics (for example, students' expectations) remain uncontrolled for in the benchmarking process, and we explore those further in Section 3.3.4

3.3.4 Grouping by provider type: a discussion

Finally, within Section 3.3 on benchmarking, we discuss how aspects not included in benchmark but that otherwise one might want to control for, could be included in the TEF assessment process. In particular, we consider the idea of conducting TEF assessments independently on groups of providers that have been specified and identified in advance of benchmarking, according to some typology or data-driven rules. We note that conducting TEF assessments by group was also one of the suggestions made by the Royal Statistical Society in their submission to the independent review of TEF (RSS, 2019, Section B) with an aim of improving comparability.

At the heart of the idea of conducting TEF assessments on groups of providers is benchmarking itself, and it's worth noting that, other than how the final awards are communicated, this is not something that would affect non-benchmarked TEF measures. The concern is that, despite benchmarking trying to 'level the playing field' by accounting for the mix of student characteristics present at each provider, there remain other factors that have an influence on the metrics' reported indicator values that are neither included in the benchmarking factors nor are what TEF is seeking to measure. As such, this section considers what benchmarking does not include or account for, in contrast to the previous subsections in Section 3.3 which consider what it does achieve.

It is difficult to define precisely what those omitted factors are (or may be) because many are difficult or impossible to measure or quantify. Some may also fail to meet the TEF benchmarking principles, for example those that are in the control of the provider (at least to some extent), and therefore should not be factored out when it comes to making comparisons.

Most factors of the sort we consider exist at provider-level (noting that benchmarking factors are applied at student-level in the current approach), and might be expected to influence the observed indicators. An example we have encountered anecdotally is student expectations, and the effect these might have on the NSS metrics. Suppose two providers, possibly found in the same location, provide all-but-identical teaching of the same subject, but one provider is a long-established and well-regarded university, whereas the other is a newer provider that doesn't have the same reputation. It might be that students expect better teaching from the former, and as a result provide less positive feedback in their reflections of the teaching, despite the teaching being the same at both. (For completeness, we note that NSS, and TEF more generally, does not assess teaching excellence directly, rather it seeks students' reflections upon it.)

Factors relating to location – cost of living in the area, local career prospects, or other factors outside providers' control that affect students' social or economic well-being – may also affect reported metrics in a similar way. Likewise, providers have a variety of missions, target-student groups, and general approaches to teaching. These latter categories start to overlap with aspects that are in the control of providers, and arguably should remain as part of the residual, d, as they are part of what TEF is aiming to capture. The benchmarking principles are clearly relevant in deciding what should be controlled for and what should not.

Inclusion of factors such as student expectations or local-area variations in the current benchmarking approach would be challenging as not all are measurable, or occur only at provider- rather than student-level. Of course, if these factors were highly correlated with those already included in benchmarking, then the consequence of not taking account of them would be small, but the extent of any correlation is unknown. An alternative approach to account better for these factors is to group providers according to their type (with 'type' to be defined), and to carry out TEF assessments separately and independently within each group.

If one were to believe that fundamental differences of the sort described exist, then assessing all providers together (as is currently the case) would lead to the following areas of possible statistical concern:

- providers of different types contributing to each others' benchmarks, and thus influencing the TEF outcomes when perhaps that is inappropriate
- the same TEF outcome for providers of different types not necessarily carrying the same meaning, as the outcome awarded would be partly a reflection on some fundamental differences between providers outside what TEF is attempting to measure; the same TEF award category would therefore not be fully comparable between providers.

Assessing independently by provider-type groups therefore has the potential to improve comparability and to 'level the playing field' further, and beyond what the current benchmarking factors offer. However, the difficulty would lie in establishing a meaningful grouping of providers: one that is associated with the additional factors we wish to account for, is recognised by students and accepted by most providers. Establishing such a grouping would not be an easy task, and its implementation would, no doubt, contain further, practical challenges.

Discussion of provider types continues in later sections of this evaluation. Sections 10.1 and 10.2 provide evidence suggesting that different provider types may receive different treatment in the TEF process anyway. Included within that is the effect of the TEF assessors' and panel's holistic judgement on the Step 1a starting points in combination with other information; that assessment could provide some mitigation for the other factors mentioned, but is also subjective.

Our discussion naturally focusses on the statistical aspects of the process. TEF also has policy aims (BIS, 2015 and 2016b, for example), such as to reward good teaching practices, increase competition between providers and improve information for students, which we recognise are agnostic of provider type. These would need to be borne in mind and balanced against the statistical view when considering the option for grouping providers in TEF (as stated previously, policy considerations are beyond the scope of our statistical evaluation).

We recommend that further research be undertaken on assessing providers separately by type, and that serious and careful consideration is given to implementing that if the research outcomes can establish an acceptable, meaningful and practical way of doing so. Annex A discusses possible approaches to determining provider groupings, as well as giving options to assist that assessment. The research itself lies outside the scope of this evaluation, but we suggest it is conducted without delay, as any decision to implement, or not, groupings by provider types would have consequences requiring further decisions on the TEF process, so an early decision would be useful.

Recommendation 7: Research further the grouping of providers by type in the TEF assessment process to account for factors that cannot easily be included in the benchmarking process with the aim of improving comparability further. If the research outcomes suggest groupings that are viable, carefully consider the options, implications and practicality of implementation.

3.4 z-scores and assumptions of normality

3.4.1 Basics

The use of z-scores in TEF is analogous to conducting a hypothesis test on the difference, d, between the indicator and its benchmark. In effect, the z-score is the test statistic for a null hypothesis of H_0: $\mu_d = 0$ (for a given provider) as it is of the form $z = (d - \mu_d \,|H_0) \,/\, \text{std}(d)$, where μ_d is the (unknown) true difference for that provider. It should be noted that the term 'standard deviation' used here actually relates to the variability of an estimated statistic rather than of a sample of data points; in statistical terms, it would more commonly be referred to as a standard error, as noted by the Department for Education (DfE, 2017a, page 42).

3.4.2 Target populations

Similar to our discussion in Section 3.2.1, the approach used assumes some super-population of students at (or graduates from) the provider, of which the actual student population or sample at any given time is just one possible realisation. This is a reasonable, conceptual paradigm in which to ground the process, given the predictive manner in which TEF is intended to be used (that is, students will want to make inferences and extrapolate from the TEF metrics calculated on the observed population in any given academic year; presumably the experiences of the exact students that passed through the higher education system in the given academic year, that is the finite population, is of lesser interest). However, the target population (the hypothetical super-population rather than the observed finite population) does not appear to be explicitly stated in the published TEF user guidance, and we recommend that this is made more explicit.

Recommendation 8: A fuller description of the target population, and any assumptions made about it, are made more explicit in the TEF user guidance.

3.4.3 Normality assumptions

An assumption is made of normality of d, the difference between the indicator p and the benchmark E for a given provider and a given metric. Although p and E are themselves likely to be non-normally distributed, we note it is only their difference that is of relevance to the assumption of normality. The assumption of normality of d is difficult to test empirically, as we have only the one observed difference, d, at any given time for any given provider: different providers each have their own distributions of d, so different providers' data should not be compared, and it is possible that the distributions change over time.

A simulation under the null hypothesis of $d = 0$ and various global values of p was conducted by Draper and Gittoes (2004). The authors concluded that the assumption of normality was reasonable, albeit with a small but positive and statistically significant mean; however, this result was based on just one assumed value of p (0.9) and just one metric (student progression). Analysis of a similar nature, which we will return to later in this section, has subsequently been performed by OfS. This broader analysis was based on six TEF core metrics for both full- and part-time modes of study, and was conducted at both provider- and a subject-level. The results again provide little evidence against the assumption of normality. Given that a simulation to assess normality requires student-level data, to which we as reviewers have not had access, and that the assumption of normality seems reasonable, we have not tested or considered that assumption further.

However, if there were any notable deviation from the assumption of normality, it could have implications on interpretation. The first is that the quoted critical values (such as $z = 1.65$, 1.96 or 2.58 corresponding respectively to 10%, 5% and 1% significance in two-tailed tests) should be

treated with some caution as they may not accurately test the significance level prescribed. That these critical values – somewhat arbitrary in nature, but chosen with reference to the standard normal distribution – result in binary-outcome categorisations is also a relevant consideration, and is discussed further in Section 3.7.

3.4.4 Estimation of the standard deviation of *d*

A closed-form expression for the estimator of the standard deviation (standard error) of *d* (used in the denominator of the calculation of the z-score) is derived analytically, as outlined in Draper and Gittoes (2004). For a given provider and a given metric, *d* can be expressed as a weighted sum of the indicator values *p* across all the cells in a cross-tabulation of providers and benchmarking groups, where the weights (written as λ in HESA's benchmarking description, and treated as constants) are a function of various sample sizes from within the benchmarking grid. Under the 'super-population' approach, treating the λ as constants disregards a source of sampling variation: although student numbers by subject and by provider may remain (relatively) constant from year-to-year, the numbers in each particular benchmark group (those defined by various combinations of age, sex, disability, social disadvantage and so on) are not.

The algebraic approach means that there is no need to estimate explicitly the standard deviation of the benchmark (which is itself a random variable, rather than fixed quantity) or the covariance between the indicator and its benchmark (which may be positive, depending on the extent to which a provider contributes to its own benchmark). However, it does rely on the assumption that the sample sizes in each of the cells in the grid are fixed rather than random, noted above. Some of these features only become clear upon inspection of the detailed methodology, so we recommend that they are made explicit in a more accessible manner in the publicly available TEF documentation.

Recommendation 9: Make more explicit in the publicly available TEF documentation details of assumptions made in the calculations of z-scores and their standard deviations.

Estimation of the standard deviation of *d* also involves shrinkage, which is a standard statistical technique for combining estimates with other sources of information. If we consider the multi-way cross-tabulation of benchmarking factors and providers (that is, a 'grid' consisting of 'cells'), shrinkage results in cell-level estimates of *p* being pulled towards their global (full-sample) mean, because of the possibility of sparseness in the grid and values of *p* close to 1.

As discussed in Draper and Gittoes (2004), the shrinkage parameter was optimised on the basis of having 5% of absolute z-scores in excess of 1.96 in null simulations, with a value of 0.5 (corresponding to the arithmetic mean of global and cell-level values of *p*) ultimately being selected. This seems a reasonable approach and choice of objective function, given the nature of the problem at hand. More recent analysis performed by OfS has shown that a shrinkage parameter of 0.5 still appears to be a reasonable choice, or at least not notably worse than anything else if a one-size-fits-all approach is to be adopted. However, there appears to be scope for allowing the shrinkage parameter to vary by metric, mode of study, and provider headcount band. For example, the percentage of absolute z-scores greater than 1.96 is notably in excess of 5% for the non-continuation metric compared with the other metrics, and it also appears to be negatively correlated with provider headcount (this is, as the student headcount-size increases, the percentage of absolute z-scores exceeding 1.96 decreases). Such analysis of the current calibration of the TEF flagging system is welcomed, and we encourage OfS to re-perform such analysis on a periodic basis.

Recommendation 10: Calibration of the TEF flagging system should be reviewed periodically, on an ongoing basis, and corrective action taken where necessary.

3.4.5 Associations between flags and size of provider

In this section we first consider the probability of a single metric of a provider being flagged, and then consider the probabilities of a provider achieving particular Step 1a outcomes (Gold, Silver, Bronze), which is based on a combination of metrics, realised flags and weights.

Forster (no date) considers the theoretical probability of a single, given metric being flagged for a range of parameters and situations, and under a number of simplifying assumptions, which do not detract from the analysis. The analysis identifies two conditions under which providers are less likely to be flagged:

- those with smaller student numbers
- those that have a large number of students relative to the total number of students.

Both observations are intuitive. Recall first that a metric is flagged if both the absolute value of the difference, d, exceeds 2.0 and likewise the z-score (defined as $z = d / \text{std}(d)$) exceeds 1.96 (the significance test). That smaller student numbers result in a lower probability of flagging relates to the significance test: a smaller sample size makes a significant result less likely (analogous to wider confidence intervals). Whilst this behaviour is statistically sound, and is exactly as hypothesis tests are designed to work, it has the potential for undesirable consequences: it inherently puts smaller providers at a disadvantage in terms of their ability to achieve a Gold starting point at Step 1a, and analogously, puts them at an advantage in terms of their relatively lower propensity to be awarded Bronze.

In a similar way, a provider that has a relatively large number of students in a finite population (as in benchmark groups) can be thought of contributing a lot towards its own benchmark; as that proportion gets larger, it becomes more difficult for the provider to be flagged on the materiality test (its indicator becomes its own benchmark in the limit). (We note that the materiality test is not actually applied if the benchmark exceeds 0.97 (DfE, 2017a, paragraph 5.60)).

Naturally, there is variation between providers regarding their student numbers (absolute and relative), but any issues about flagging-probabilities will be more acute at subject-level than overall because of the smaller samples (student numbers) involved at subject level, and that fewer providers might teach particular subjects so the relative contribution would be larger.

The graphs shown in Forster (no date, Figure 1) show some low probabilities of being (negatively) flagged, even for providers that are truly (super-population) below the benchmark (and more so than the material threshold of 2 percentage points) despite moderate sample sizes. For example (Figure 1, bottom-left diagram of Forster (no date)), a provider with student size of about 300 (within a given subject domain), and with a true difference of 5 percentage points below a benchmark of 80% has only a 0.5 probability of being negatively flagged; it is just as likely as not to be negatively flagged. For a provider with a true difference of 10 percentage points below a benchmark of 80%, the probability of being negatively flagged falls to 0.5 for headcounts as low as about 84 students (Figure 1, bottom-right diagram of Forster (no date)).

These results were derived analytically by Forster (no date) according to the theoretical properties of TEF, but they potentially have real implications for the observed TEF data (Table 4a). For example, for the TEF Year Four provider-level core metrics, 4.7% of provider-metric combinations had a denominator less than or equal to 300 students and an observed difference of at least 5 percentage points below a benchmark of not more than 80%. Within smaller domains in TEF Year Four, this figure increased to:

- 6.4% for the provider-level split metrics
- 17.0% for the pilot subject-level core metrics
- 37.5% for the pilot subject-level split metrics.

It should be noted that our analysis is contingent on observed sample statistics rather than hypothetical superpopulation parameters (which are, of course, unknown). For example, the analysis reported by Forster (no date) is stratified according to true (superpopulation) differences from each provider's benchmark; we are only able to work with the published sample estimates of *d*.

Table 4a. Units* with a benchmark of ≤ 80% that are susceptible to having a probability of being negatively flagged of less than 0.5 (selected scenarios), TEF Year Four provider- and pilot subject-level data

Level	Core metrics		Split metrics	
	Difference ≤ -5 & denominator ≤ 300	Difference ≤ -10 & denominator ≤ 80	Difference ≤ -5 & denominator ≤ 300	Difference ≤ -10 & denominator ≤ 80
Provider	101 (4.7%)	28 (1.3%)	2,171 (6.4%)	704 (2.1%)
Subject	5,269 (17.0%)	3,173 (10.3%)	230,872 (37.5%)	192,645 (31.3%)

** A unit is a provider-metric combination for provider-level data and a provider-subject-metric combination for subject-level data*

The graphs presented in Forster (no date, Figure 2) also suggest that providers contributing more than about 50% to their own benchmark (defined here as the student size at the provider divided by student size overall) may start to see a reduced probability of being flagged (depending on assumptions and conditions), and those contributing more than about 80% to their own benchmark are much more likely to see a smaller probability of a given metric being flagged.

At provider-level, self-benchmarking is unlikely to be particularly problematic in the observed TEF Year Four data as the providers' student sizes are all small relative to the total student numbers (Table 4b). However, for subject-level core metrics, 6.5% of provider-subject-metric combinations that contributed at least 60% to their own benchmark had an observed difference of at least 5 percentage points below their benchmark, increasing to 26.7% for subject-level split metrics.

Table 4b. Units* that are susceptible to having a probability of being negatively flagged of less than 0.5 because of self-benchmarking (selected scenarios), TEF Year Four provider- and pilot subject-level data

Level	Core metrics		Split metrics	
	Difference ≤ -5 & contribution to benchmark# ≥ 60%	Difference ≤ -10 & contribution to benchmark# ≥ 80%	Difference ≤ -5 & contribution to benchmark# ≥ 60%	Difference ≤ -10 & contribution to benchmark# ≥ 80%
Provider	1 (0.0%)	0 (0.0%)	19 (0.1%)	0 (0.0%)
Subject	2,018 (6.5%)	2,016 (6.5%)	164,185 (26.7%)	164,001 (26.6%)

** A unit is a provider-metric combination for provider-level data and a provider-subject-metric combination for subject-level data*
 # As quoted in published TEF statistics. This is a weighted average of the provider's student size to the overall (total) student size in the domain of interest (for example, subject)

A single metric being flagged (or not) is, of course, not the same as a provider's starting point category, and the probability of being flagged in single metric does not translate simply to the probability of receiving a Gold, Silver or Bronze starting point. Rather, the Step 1a starting-point calculation depends on the realised flags for a set of six or nine metrics (not all of which are independent, as discussed in Section 3.5) in combination with their flag values (metric weights).

That complexity makes the theoretical estimation of the probability of gaining any given Step 1a starting-point category rather complicated. An alternative possibility to estimating these may be a simulation-based approach under a set of simplifying assumptions; we have not explored either option further. However, Section 3.7 contains further discussion on 'borderline' cases, which we define as metrics that are either 'just flagged' or 'just not flagged', and the implications such cases could have.

We have also performed our own analysis of associations between headcount and Step 1a starting points using the observed TEF data, leading to results that are somewhat less clear-cut than those outlined above for a single metric. For each of TEF Years Three and Four, we estimated a multinomial logistic model by regressing each provider's derived Step 1a starting point (using Silver as the reference category) on their student headcount (see Table 5). We found statistically significant associations between the odds of achieving Gold or Bronze at Step 1a and student headcount in TEF Year Four, with the odds of achieving Gold or Bronze rather than Silver increasing with headcount. The results for TEF Year Three were not significant, possibly reflecting the smaller number of providers than in Year Four.

We analysed the subject-level 2018/19 TEF pilot data in a manner largely analogous to that for the provider-level data as outlined above, except that: we were not supplied headcount data for the subject-level data, so we instead estimated the association between Step 1a starting point and *denominator size* (averaged across the nine metrics for each provider-subject combination, using metric weights prescribed under the 2018 method); and we included random intercepts to account for intra-provider and -subject correlation (which necessitated the use of binary rather than multinomial logistic regression). As with the provider-level TEF Year Four data, we found statistically significant associations between the odds of achieving Gold or Bronze at Step 1a and the denominator size at the subject level (Table 6), with the odds of achieving Gold or Bronze rather than Silver increasing with denominator size.

The analyses summarised in Tables 5 and 6 are intended to be descriptive in nature and are restricted to estimating only bivariate associations between the Step 1a starting point and student headcount / denominator size; our intention is not to construct a model that explains or predicts the Step 1a starting point in some broader sense, so a full set of model diagnostics is not reported here. In particular, it should be noted that we have not controlled for factors that could be confounded with headcount (such as provider type); identifying such factors is beyond the scope of this evaluation. Moreover, larger providers may contribute more to their own benchmark (self-benchmarking) and therefore, all else being equal, the decrease in the standard deviation of *d* (the denominator of the z-score) will be offset by a decrease in the value of *d* itself (the numerator of the z-score). The analysis conducted by Forster (no date) supports the preceding logic that self-benchmarking reduces providers' propensity for being flagged, and this effect will be inherent in our analysis but has not been controlled for.

Table 5. Output of multinomial logistic regression associating Step 1a starting point with student headcount, provider-level TEF Years Three and Four

TEF Year	Step 1a starting point	Odds ratio (Student headcount)	P-value
TEF Year Three	Gold	0.99998	0.8503
	Silver	-	-
	Bronze	1.00003	0.5315
TEF Year Four	Gold	1.00004	0.0209
	Silver	-	-
	Bronze	1.00004	0.0151

Note: odds ratios estimated on 86 observations for TEF Year Three and 396 observations for TEF Year Four

Table 6. Output of binary logistic regression associating Step 1a starting point with denominator size, subject-level 2018/19 TEF pilot

Step 1a starting point	Odds ratio (No. reportable metrics)	P-value
Gold	1.00104	0.0001
Silver	-	-
Bronze	1.00141	< 0.0001

Note: odds ratio for Gold estimated on 1,994 provider-subject combinations with a Gold or Silver starting point at Step 1a; odds ratio for Bronze estimated on 2,274 provider-subject combinations with a Bronze or Silver starting point at Step 1a

The odds ratios presented in Tables 5 and 6 above should be interpreted as the proportional change in the odds of obtaining Gold/Bronze (rather than Silver) at Step 1a associated with a *unit* increase in the number of students at the provider; the estimated ratios may be small in magnitude, but they are on a *per-student* basis. Tables 7 and 8 therefore consider the odds ratios associated with achieving Gold/Bronze for providers with various student headcounts / denominator sizes compared to one with the median student headcount / denominator size. For example, for provider-level TEF Year Four (Table 7), the 90th percentile of headcount was 13,130 students; the odds of achieving Gold or Bronze for a provider with this many students were 1.696 or 1.576 times greater, respectively, than those for a provider with the median headcount of 588 students. At the 99th percentile (23,811 students), the odds ratios increased to 2.659 for Gold and 2.321 for Bronze. Similar findings were observed at subject-level (Table 8) but, given the larger regression coefficients, the odds ratios were more pronounced than those at provider-level.

Table 7. Odds ratios for achieving Gold/Bronze at Step 1a for various student headcounts vs. median student headcount, provider-level TEF Year Four

Headcount (quantile)	Headcount (no. students)	Odds ratio for Gold Step 1a starting point (vs. median headcount)	Odds ratio for Bronze Step 1a starting point (vs. median headcount)
1st percentile	25	0.977	0.980
10th percentile	95	0.980	0.982
25th percentile	224	0.985	0.987
Median *(reference)*	588	1.000	1.000
75th percentile	3,363	1.124	1.106
90th percentile	13,130	1.696	1.576
99th percentile	23,811	2.659	2.321

Table 8. Odds ratios for achieving Gold/Bronze at Step 1a for various denominator sizes vs. median denominator size, subject-level 2018/19 TEF pilot

Denominator (quantile)	Denominator (no. students)	Odds ratio for Gold Step 1a starting point (vs. median denominator)	Odds ratio for Bronze Step 1a starting point (vs. median denominator)
1st percentile	15	0.882	0.843
10th percentile	32	0.898	0.864
25th percentile	55	0.919	0.892
Median *(reference)*	136	1.000	1.000
75th percentile	313	1.202	1.283
90th percentile	578	1.582	1.864
99th percentile	1,652	4.819	8.459

3.4.6 Use of z-scores

Multiple comparisons

We note first that two providers with identical z-scores but based upon very different sample sizes are not necessarily directly comparable: the interpretations of those z-scores should differ (as noted in Draper and Gittoes (2004)). Their approach suggests visually plotting $d_j \pm 1.96 \times \text{std}(d_j)$ for all providers, j, arranged in rank-order of d and noting which confidence intervals include zero. The thresholds used to denote 'meaningful' differences (currently $d = \pm 2.0$) can easily be added too. This is clearly a useful suggestion, and we recommend that plots such as these are made available to TEF assessors and the TEF Panel. Indeed, we present an example of such a plot using the published TEF Year Three metrics data later in this section (see Figure 3).

Recommendation 11: Improve communication on statistical uncertainty. For example, publish plots of TEF-metric differences and their confidence intervals by provider in rank order of the differences. The plots would clearly show which confidence intervals include zero, and which differences have absolute values that exceed thresholds considered to be meaningful.

We now consider the more general case of multiple comparisons. When interpreting the z-scores, there is a risk associated with making multiple comparisons, in which analysts (prospective students or a TEF Panel, for example) consider and compare the z-scores of numerous providers at once. In such a scenario, there becomes a much larger probability of observing a 'significant' result by chance alone than the intended 5% (for example) when there is, in fact, no real difference.

When making multiple comparisons, the analyst may choose to control the family-wise error rate (the probability of experiencing one or more false-positive findings) or the false discovery rate (the proportion of findings that are false-positives). A common practice is to inflate the critical value (and there are various methods for doing this), but the extent of this depends upon the number of comparisons made. This cannot be known in advance when the users and uses are not specified and many. We suggest, at least, that appropriate wording is added to guidance about use of z-scores when comparing providers, and it may be useful to provide refresher training about the topic to assessors and members of the TEF Panel.

Recommendation 12: Appropriate guidance on making multiple comparisons should be drafted and made prominent in the user guidance and with TEF outputs. The recommended plots of differences and confidence intervals could also accommodate this, with further extensions to the intervals' widths for multiple-comparison purposes also shown.

We have used the publicly available TEF Years Three and Four data to test the impact of correcting for multiple comparisons, and the effect seems likely to be non-negligible. We begin with the following set of assumptions:

- the number of tests to be conducted at Step 1a is equal to the number of providers multiplied by the number of metrics, less the number of non-reportable metrics and those relating to the minority mode of the study
- for each of these tests, a null hypothesis of $d_i = 0$ is assumed
- the family-wise error rate is to be controlled at the 5% level
- the materiality test is fixed at $|d_i| \geq 2$.

In TEF Year Three, 463 tests are to be performed for the core metrics in total, resulting in a Bonferroni-corrected critical z-score of ±3.87 (we note that we have not considered any pairwise comparisons of providers in this assessment). The effect of applying this correction is to reduce the number of provider-metrics flagged as positive ('++' or '+') from 86 to 33, and those that are flagged as negative ('--' or '-') from 77 to 42, with the number of unflagged metrics (also called neutral flags ('=')) increasing from 300 to 388. The impact of increasing the critical z-score from ±1.96 to ±3.87 is visualised in Figure 3, where the increase in the width of the confidence intervals around estimates of *d* is marked. The effect of controlling for multiple comparisons is even more pronounced for the split metrics because of the increased number of null hypothesis tests being conducted (7,187 in TEF Year Three), which pushes the critical z-score to ±4.50: the number of provider-split-metrics flagged as positive falls from 1,012 to 248, and those that are flagged as negative from 1,084 to 344.

In TEF Year Four, 2,155 tests are to be performed for the core metrics in total, resulting in a Bonferroni-corrected critical z-score of ±4.23. The effect of applying this correction is to reduce the number of provider-metrics flagged as positive from 397 to 185, and those that are flagged as negative from 403 to 185, with the number of unflagged metrics increasing from 1,355 to 1,785. In terms of the split metrics, 34,146 tests are to be conducted in TEF Year Four, which pushes the critical z-score to ±4.82: the number of provider-split-metrics flagged as positive falls from 5,286 to 1,493, and those that are flagged as negative from 5,323 to 1,501.

Applying the existing metric weights and rules (that is, the 2017 method) results in: 9 out of 19 providers being re-categorised from Bronze to Silver and 4 out of 5 providers being re-categorised from Gold to Silver in TEF Year Three; and 58 out of 96 providers being re-categorised from Bronze to Silver and 14 out of 30 providers being re-categorised from Gold to Silver in TEF Year Four.

If the use of flagging is to continue in its current form (a topic we consider further in Section 4), then it may be prudent to increase the critical value of the z-score to mitigate against the incorrect flagging of results. However, we note that the existing double-flag options ('++' and '--') are based upon critical values of $|z| \geq 3.00$, and in that sense their use would help mitigate against erroneous conclusions being drawn when making multiple comparisons.

However, even if a threshold of $|z| \geq 3.00$ is used in the Step 1a ruleset (and all other aspects of the methodology, including the $|d| \geq 2$ threshold, remain unchanged), we note that the impact of correcting for multiple comparisons remains substantial. For the core metrics in TEF Year Three, the number of provider-metrics flagged as positive falls from 46 to 33, and those that are flagged as negative from 130 to 42; while in TEF Year Four, the number flagged as positive falls from 265 to 185, and those that are flagged as negative from 631 to 185.

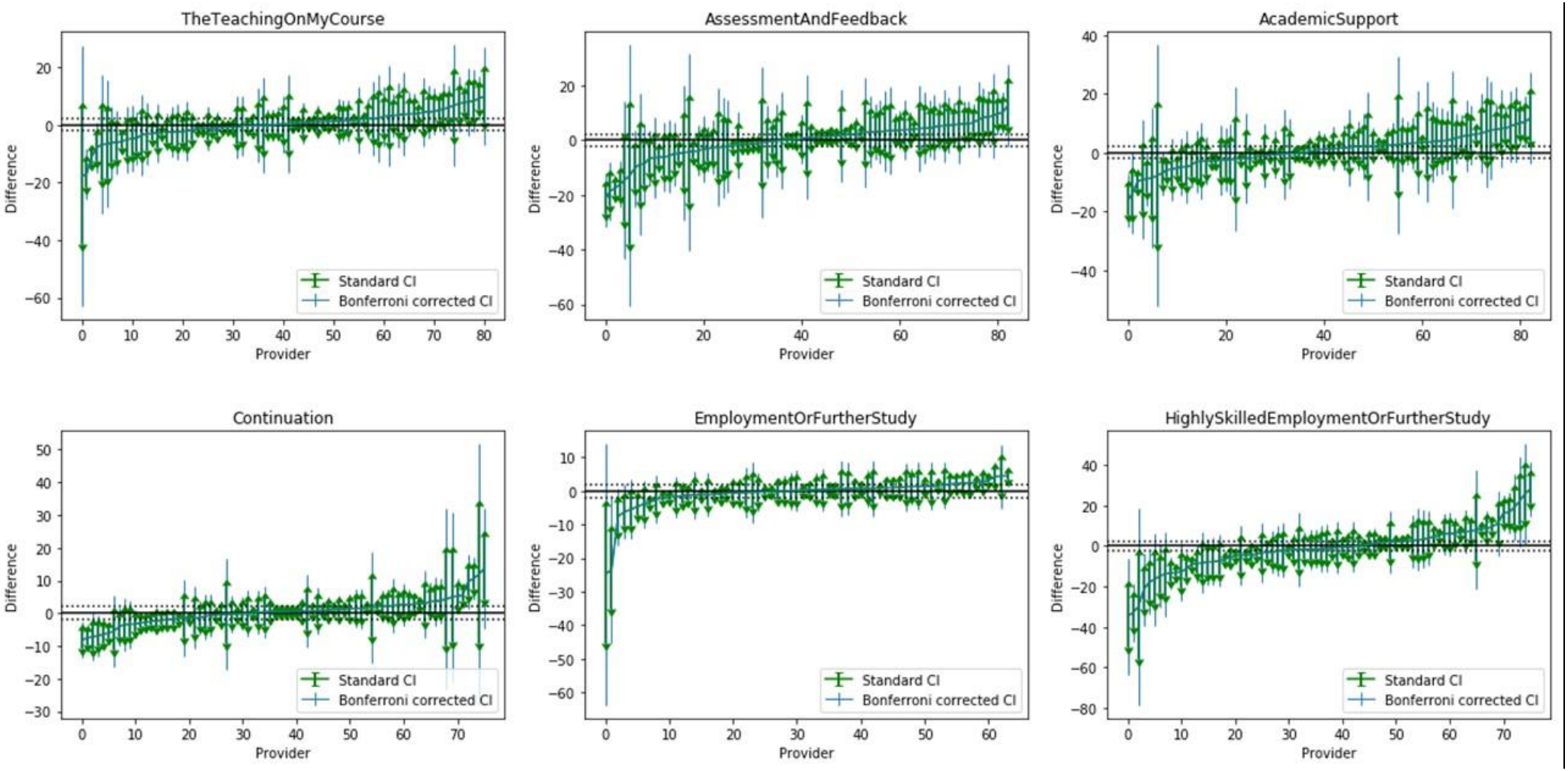
In terms of Step 1a starting points, however, single and double flags are not distinguished, thus it is the less stringent (single-flag) criteria that are used to derive the Step 1a Gold/Silver/Bronze outcomes, and it is not clear quite to what extent the TEF Panel consider the double-flags in their holistic judgement. We are aware that OfS is currently undertaking some simulation-based analyses of alternative thresholds to derive the flags in Step 1a that take into account the issue of multiple comparisons; we welcome this analysis, and we look forward to the results being published.

When evaluating the impact of correcting for multiple comparisons on the TEF Year Four data, it should be remembered that metrics data and the resulting flags are published for all providers in the scope of TEF in England, irrespective of whether they were assessed for a TEF award, plus those from the devolved administrations that chose to participate. This is in contrast with TEF Year Three, where the published metrics data relate only to those 86 providers that were assessed. The result is that the Bonferroni-corrected critical z-score is larger in absolute terms in TEF Year Four than it is in TEF Year Three, because of the greater number of hypothesis tests being performed. It should be noted that the TEF assessors and Panel may only require tests to be performed for the providers under assessment in any given year, hence reducing the impact of correcting for multiple comparisons. However, users of the published metrics data, including students, will see and potentially use all of the information available to them in a given year, so the number of tests to be corrected for is not necessarily just those that involve the providers undergoing TEF assessment.

We acknowledge that the Bonferroni correction is known to be overly conservative. Indeed, other approaches to controlling the family-wise error rate such as the Holm method (Holm, 1979) are often preferred as they reduce the loss of power in the test for a given type-I error rate. Our intention here is not prescribe a preferred method, but rather to illustrate the potential impact on statistical inferences of applying such a correction to the TEF metrics.

We re-emphasise here that a particular difficulty with calculating corrected z-score thresholds is that the number of tests to be conducted is generally unknown *a priori* without making some assumptions. For example, if the TEF Panel simply want to test each provider-metric's observed difference against a fixed quantity such as $d = 0$ (as is implied by the current z-scores), then the number of comparisons is equal to the number of providers being assessed multiplied by the number of metrics (less any non-reportable metrics). However, what if a student wishes to use the published TEF metrics data to compare performance *between* several providers? In this situation, the number of comparisons to be made very much depends on the list of providers that the student is interested in. In the most extreme situation, the number of tests to be conducted is equal to the total number of possible combinations between every pair of providers (for example, every combination of two providers drawn from the assessed population of 86 providers for TEF Year Three) – which has the potential to increase the z-score threshold to a sizeable value.

Figure 3. 95% confidence intervals around estimates of *d* before and after applying the Bonferroni correction, stratified by core metric, TEF Year Three



Concluding remarks on z-scores

In terms of the TEF Step 1a process, the set of z-scores is one input to the rules used to determine the categorisation of providers (and subjects), and one which could be used for the purposes of ranking providers (though this is not the purpose of TEF at present). In that sense, any moderate discrepancies in assumptions may be of lesser importance than other aspects that follow (such as the TEF assessors' and Panel's evaluation of the contextual data and providers' written submissions). It is very difficult to know the effects of deviations from assumptions, especially with each provider (and subject) having its own distribution of the difference *d*, and it is difficult to evaluate quantitatively in light of the later stages of the TEF assessment process.

3.5 Correlation between core metrics

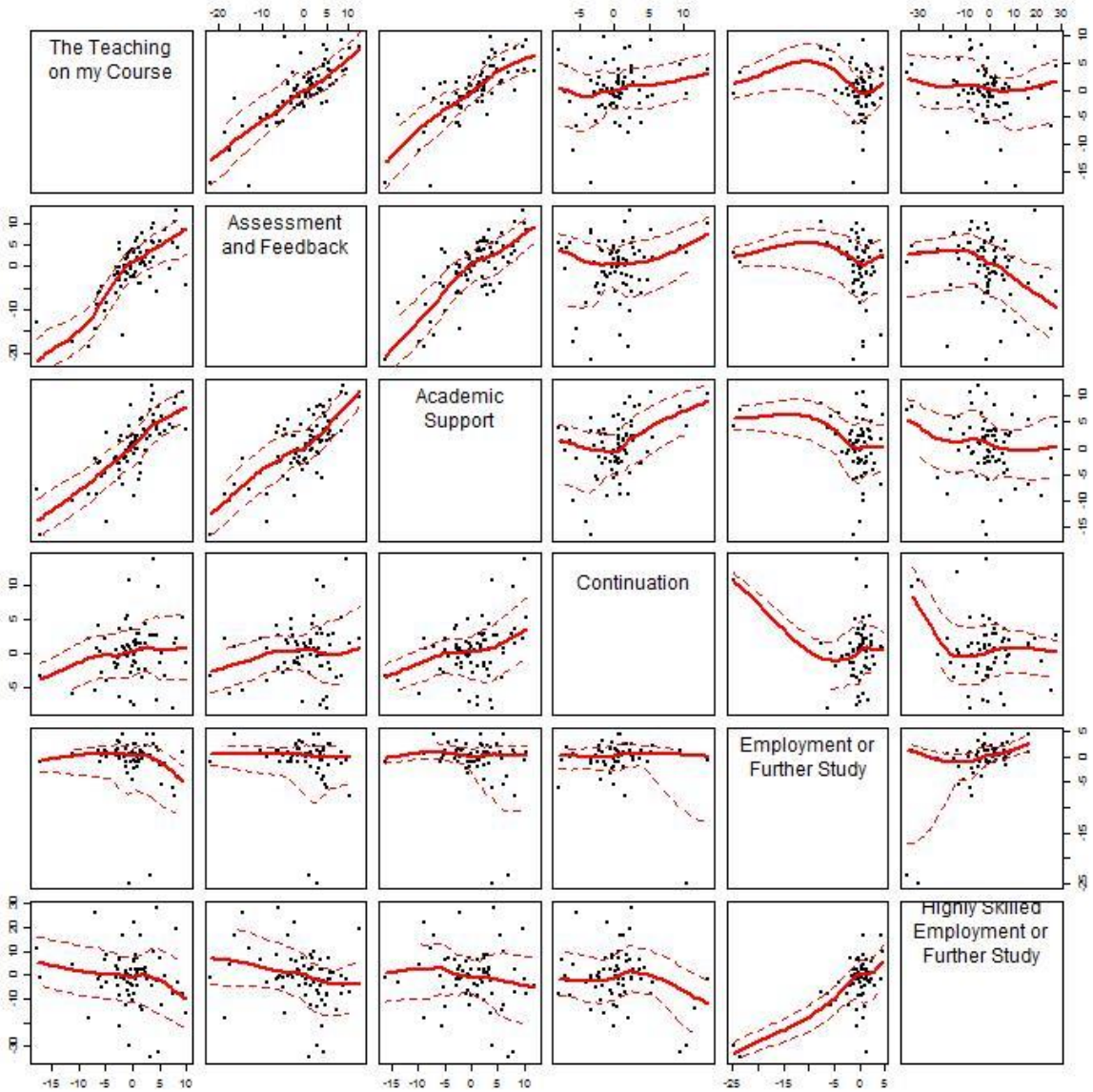
We have assessed the empirical correlation between each pair of core metrics for TEF Years Three and Four (excluding non-reportable and minority-mode metrics). Figures 4a and 4b illustrate that there is a high degree of positive linear correlation between the three NSS metrics, but less so between the other metrics (including between NSS and non-NSS metrics). In TEF Year Three, the Spearman correlation coefficients between pairs of NSS metrics (see Table 9) were: 0.792 between 'Academic support' and 'The teaching on my course'; 0.676 between 'Assessment and feedback' and 'The teaching on my course'; and 0.666 between 'Academic support' and 'Assessment and feedback'. In TEF Year Four, the corresponding coefficients were: 0.822 between 'Academic support' and 'The teaching on my course'; 0.682 between 'Assessment and feedback' and 'The teaching on my course'; and 0.728 between 'Academic support' and 'Assessment and feedback'. The association patterns between all other pairs of metrics were far less pronounced in both years, and the correlation coefficients were far smaller.

These results suggest a lack of independence between the NSS metrics and, empirically at least, that these collinear metrics may be accounting for the same variation in what TEF is trying to capture. We note that the NSS metrics individually carry a smaller weight than each of the other metrics (under both the 2017 and 2018 methods), which to some extent may mitigate the impact of the collinearity on the Step 1a starting point. DfE (2017b) shows that some consideration of the collinearity between metrics in TEF Year Two had taken place, and resulted in the weights assigned to the NSS metrics being reduced in Year Three. It should be noted that the correlation coefficients between pairs of NSS metrics reported by DfE (2017b) are markedly higher than those summarised in Table 9: 0.996 between 'Academic support' and 'The teaching on my course'; 0.946 between 'Assessment and feedback' and 'The teaching on my course'; and 0.972 between 'Academic support' and 'Assessment and feedback'. The precise method by which these correlation coefficients were calculated is not clear from DfE's published note.

An alternative approach to reducing the weights on the NSS metrics in a fairly arbitrary way could be the use of a technique such as principal components analysis or factor analysis; methods such as these are in line with international guidance for constructing weights for composite indicators (OECD, 2008). Stated-preference techniques present another possibility, whereby students would be surveyed to elicit their priorities in terms of educational outcomes. These approaches may be worth consideration, but their implementation could introduce unnecessary complexity to the methods, and may be difficult for non-statistical users to understand. In contrast, the current approach of weights shows quite transparently how each metric contributes. As such, we do not make any formal recommendation to adopt such approaches, but their use could form part of any analysis used to determine appropriate weights for the metrics.

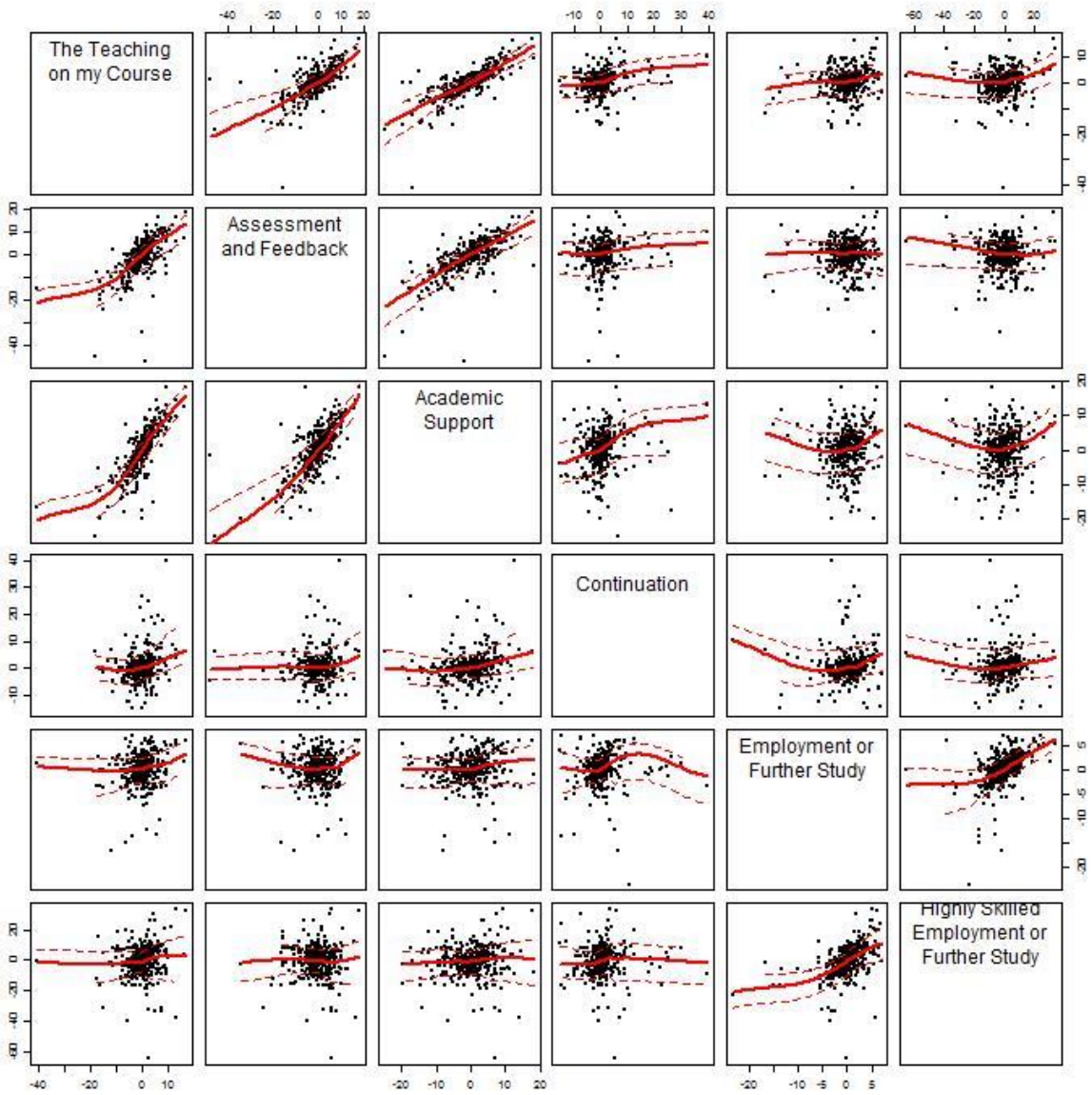
Finally, we reiterate that our analysis of correlation (dependence) between metrics is entirely empirical, based on the observed metrics data. We have not assessed the degree to which the metrics are actually measuring the same concepts; given that our area of expertise is the methodology associated with official statistics, such an assessment would be better left to institutions that specialise in education policy research.

Figure 4a. Smoothed scatterplot matrix, core metrics, TEF Year Three



Note: smoothed lines estimated by LOESS; confidence intervals are at the 95% level

Figure 4b. Smoothed scatterplot matrix, core metrics, TEF Year Four



Note: smoothed lines estimated by LOESS; confidence intervals are at the 95% level

Table 9. Pairwise Spearman correlation coefficients, core metrics, TEF Years Three (top numbers) and Four (bottom numbers)

	The teaching on my course	Assessment and feedback	Academic support	Continuation	Employment or further study	Highly skilled employment or further study
The teaching on my course	1.000 1.000	0.676 0.682	0.792 0.822	0.195 0.238	-0.179 0.149	-0.137 0.143
Assessment and feedback		1.000 1.000	0.666 0.728	0.085 0.089	-0.215 0.011	-0.328 -0.070
Academic support			1.000 1.000	0.315 0.270	-0.158 0.138	-0.163 0.103
Continuation				1.000 1.000	0.173 0.188	0.017 0.136
Employment or further study					1.000 1.000	0.362 0.566
Highly skilled employment or further study						1.000 1.000

3.6 Comparison of weights between the 2017 and 2018 methods

Tables 2a and 2b facilitate a comparison of the weights ascribed to the core metrics between the 2017 and 2018 methods, respectively. We now consider this further.

The weights are presented in the DfE and OfS documentation (DfE, 2017a; OfS, 2018.44) as absolute values, and their totals differ (4.5 compared in the 2017 method, compared with 7.5 in the 2018 method). Having different totals makes comparison of the weights somewhat more difficult, and is the reason we calculated relative weights (parts per 100) for both. Indeed, unless the reader is careful, use of only the absolute weights may even be misleading.

As an example, the 'Continuation' metric's absolute weight has increased from 1.0 to 2.0 between the 2017 and 2018 methods, which may give the impression that it is now twice as important. In fact, its relative contribution has increased only from 22.2% to 26.7%: an increase by a factor 1.2, much smaller than it might have otherwise appeared.

Recommendation 13: To improve transparency, TEF should adopt use of relative weights for the core metrics, rather than absolute weights.

OfS (2018.44) states that the weights were determined by DfE and briefly outlines some objectives, but does not detail how these weights were decided, and we recommend that further detail is published. Our interpretation, from OfS (2018.44) and from DfE (2017b), is that the six core metrics originally had equal weights, with those for the NSS metrics then being halved to acknowledge collinearity, and the weight for the 'Continuation' metric subsequently being increased for the Year Four pilot; at the same time, other metrics were introduced. We would recommend that an expert panel should discuss and decide on the relative merits of the metrics, perhaps based on evidence from a consultation of user views, as there is no one 'correct' answer: the choice of weights is at least as much a question of education policy and the intended outcomes of higher education as it is a statistical one.

Recommendation 14: Convene an expert panel to decide on the metrics' weights (if not done already), and clearly communicate the principles, decisions, and the rationale for them.

Where empirical techniques are adopted, they should be conducted in accordance with international guidance and best practice on deriving composite indicators from weighted input variables; for example, see material published by the OECD (2008). Section 4.2 of the present report considers related topics on composite indicators.

Naturally, different users would have their own views on which metrics are most important and which less so. A prospective student, for example, may be primarily interested in comparing different providers only on their teaching quality and learning environment, for example, and not be so concerned about what other, previous students have achieved after graduation.

There should, of course, be a single, authoritative set of weights from which the TEF assessment is generated, but it may be useful to develop and make available a tool for users in which they choose and input their own set of weights to allow a more meaningful, personal comparison of providers at Step 1a.

Recommendation 15: Consider developing a 'personal TEF calculator': a tool that allows users to input their own metric weights for the Step 1a calculation.

3.7 Binary nature of flags

At present, each metric can contribute to the Step 1a staring point a value of only all or nothing of its prescribed weight, depending on which side of a threshold the provider's observed values of d and z lie. Such an approach is comparable with those of the UK Performance Indicators for Higher Education Providers (HESA). However, use of a binary cut-off can produce undesirable outcomes: two providers with arbitrarily close values of z (or d) can receive starkly different starting points at Step 1a. The following example, based on the 2018 method, illustrates this.

Box 1. Hypothetical example illustrating an undesirable property of the binary nature of flags (2018 method)

Suppose two providers are identical in respect of all metrics and flags except 'Continuation'.

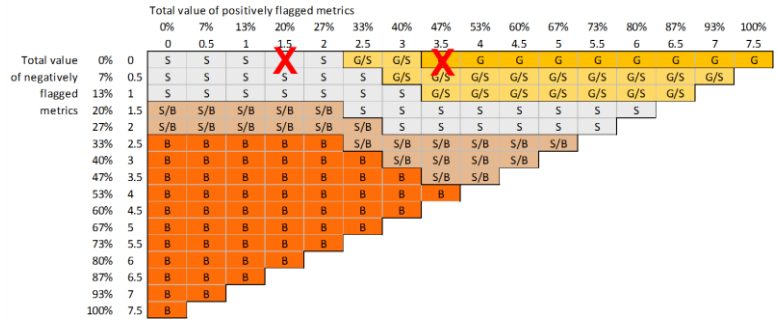
Suppose neither provider has any negative flags, and the total value of the other positive flags is 1.5. That means we are considering the very top line of the starting point diagram shown below for both providers.

Suppose now the 'Continuation' metric z-score is > 1.96 for both providers, but the difference, *d*, for the first provider is 1.99 and for the second it is 2.01.

The first provider's 'Continuation' metric remains unflagged, and gets a flag value of zero; its total positive value remains at 1.5.

The second provider's 'Continuation' metric is then flagged and receives its flag value of 2.0; its total positive value thus becomes 3.5.

Thus, the first provider's starting point at Step 1a is Silver, and the second provider's is Gold; the two providers' starting points are displayed with red 'X' symbols in the diagram below:



Not only are the providers' starting points different, but neither is in the Gold/Silver borderline category, despite their metrics being only marginally different from each other.

Statistical uncertainty, in a similar way, is ignored by the binary nature of the threshold approach: an observed difference, *d*, of 1.99 may well be observed from an underlying distribution with a true value that is 2.01 or greater (that is, on the other side of the threshold), and so on. Likewise, a value of *z* of 1.95 could easily represent a value of 1.96 or more because assumptions made about the normal distribution are not quite met.

The choice of 5% as the significance level (corresponding to the critical value of 1.96) and a material difference being defined as 2.0 percentage points are arbitrary in any case. For example, if a 'borderline difference' is defined as $1.80 \leq |d| < 2.00$ (10% lower than the current threshold of 2.00), and a 'borderline z-score' is defined as $1.88 \leq |z| < 1.96$ (corresponding to the 6% rather than the 5% significance level), then the following 'borderline' cases are present in the published TEF metrics data and represent cases that weren't flagged but 'might have been':

- **TEF Year Three core metrics:** 11 provider-metrics with borderline *d*, and eight provider-metrics with borderline *z* (but none with both)
- **TEF Year Three split metrics:** 203 provider-metric-splits with borderline *d*, and 105 provider-metric-splits with borderline *z* (but just two with both)
- **TEF Year Four core metrics:** 77 provider-metrics with borderline *d*, and 38 provider-metrics with borderline *z* (but none with both)
- **TEF Year Four split metrics:** 1,107 provider-metric-splits with borderline *d*, and 610 provider-metric-splits with borderline *z* (but just 14 with both).

We also consider borderline cases that are *above* the flagging criteria (cases that were flagged, but 'only just'), and define those here simply as meeting the single- rather than double-flagging rules for *d* and/or *z*:

- **TEF Year Three core metrics:** 73 provider-metrics with borderline *d*, and 82 provider-metrics with borderline *z* (eight with both)
- **TEF Year Three split metrics:** 975 provider-metric-splits with borderline *d*, and 1,092 provider-metric-splits with borderline *z* (106 with both)
- **TEF Year Four core metrics:** 304 provider-metrics with borderline *d*, and 357 provider-metrics with borderline *z* (46 with both)
- **TEF Year Four split metrics:** 4,770 provider-metric-splits with borderline *d*, and 5,287 provider-metric-splits with borderline *z* (730 with both).

Box 2. Case study: Illustration of closeness-to-boundaries with one particular, real example; the provider has been chosen for no reason other than having realised metrics that illustrate our discussion

Sparsholt College, TEF Year Three core metric data for full-time students (the provider's majority mode of study):

Metric	Weight	*d*	*z*	Flag
1. The teaching on my course	0.5	3.78	1.86	=
2. Assessment and feedback	0.5	5.53	2.07	+
3. Academic support	0.5	5.01	2.18	+
4. Continuation	1.0	2.54	2.11	+
5. Employment or further study	1.0	2.37	2.04	+
6. Highly skilled employment or further study	1.0	-1.67	-0.72	=

Source: https://www.officeforstudents.org.uk/advice-and-guidance/teaching/tef-outcomes/#/provider/10006050

The combination of flags realised gives this provider a Gold Step 1a starting point (total positive flag value of 3.0 and no negative flags – see Figure 1 for the 2017 method Step 1a starting-point diagram)). However, although Metrics 2 to 5 are all flagged, none is double-flagged) and some are arguably very close to not being flagged as their z-scores all lie between 2.04 and 2.18 (the boundary is $z = 1.96$), which correspond to two-tail significance levels of 4.1% to 2.9%, in the same way that $z = 1.96$ corresponds to 5.0% significance and $z = 2.58$ to 1% significance.

If just one of Metrics 4 or 5 (which have *d* values closest, and also relatively close, to the threshold of $d = 2.0$) had not been flagged, this provider's Step 1a starting point would have been Silver instead of Gold. We also note that Metric 1, although unflagged, is arguably close to being flagged.

Thus we have seen that the binary nature of flagging metrics can have undesirable consequences for the Step 1a starting-point calculations insofar as different starting-point categories can be generated from quite small differences in individual metrics. Although it would move away from practice adopted by the UK Performance Indicators for Higher Education Providers (HESA), we strongly recommend that the binary nature of achieving all or nothing of the weight, depending on arbitrary thresholds, is changed in the Step 1a calculation, so that some proportion of the weight can be obtained instead. The other main use of flags, the visual identification of significant and meaningful differences in the TEF workbooks, could continue.

In Annex B we suggest a possible approach to allocating a proportion of the weight using such a mathematical function, which we will call *prop*. Such a function could form the start of the redevelopment work required. Any new approach will require careful development and the input of specialists in the field; before adoption, thorough testing would be required and further development of other aspects, such as the starting point regions, may also be necessary.

Recommendation 16: Carefully develop, test and implement (assuming feasible) an alternative approach to the binary nature of flag values when used in the Step 1a calculation.

3.8 Missing metrics

Some metrics are not reportable (see, for example, OfS (2018.44a)), for various reasons. This section considers the prevalence and impact of non-reportable core metrics; non-reportable split metrics are covered in Section 5.2.

In TEF Year Three, the mean number of core metrics reported across the 86 assessed providers was 5.38; all providers reported at least one core metric, 91% of providers reported at least four core metrics, and 71% reported all six. In TEF Year Four, the mean number of core metrics reported across the 411 providers for which metrics data are published was 5.24; 96% of providers reported at least one core metric, 87% reported at least four, and 72% reported all six. Non-reportable core metrics therefore do not appear to be a substantial problem, at least in terms of their prevalence.

When quantifying the prevalence of non-reportable metrics in the published TEF Years Three and Four datasets, we have assumed that any provider-metric with a null-value reason code of ‘N’, ‘N/A’, ‘R’, ‘SUP’ or ‘DP’ is non-reportable, as specified in the published documentation relating to the 2017 method (DfE, 2017a). However, of the 32 provider-metrics in the TEF Year Three dataset that have a reason code of ‘DP’ (numerator differs from the denominator by fewer than three students), 17 have a ‘+’/‘++’/‘-’/‘--’ (that is non-neutral) flag, despite the data needed to calculate these flags being suppressed. It is unclear from the published TEF documentation as to whether flags such as these are considered by members of the TEF Panel, and/or whether the underlying (non-suppressed) data are made available to them.

Where a particular provider’s core metric is non-reportable, our understanding is that the current process involves ‘imputing’ a flag (rather than the indicator or the difference) based on the metric’s three component years, as long as at least one of these years is itself reportable: if at least one of the component years has a ‘+’ or ‘++’ flag, then the overall core metric is assigned the same flag; if at least one of the component years has a ‘-’ or ‘--’ flag, then the overall core metric is assigned the same flag; if none of the component years has a ‘+’, ‘++’, ‘-’ or ‘--’ flag then the overall core metric is left unreportable. The impact of this applying this imputation method to the TEF Year Three data is to increase the number of Gold starting points at Step 1a from five to six, and increase the number

of Bronze starting points from 19 to 22. (Aside: throughout this evaluation, the aforementioned imputation method has not been applied when conducting analyses that involve deriving Step 1a starting points, because details of the method do not appear to be in the public domain and were only made known to the reviewers in the latter stages of our work.)

The simple and pragmatic imputation approach outlined above raises several concerns. Firstly, the fact that the imputations favour '+'/'++' and '-'/'--' flags over unflagged (neutral or '=') ones means that the approach is biased towards significant and material differences. For example, given a situation where the core metric is non-reportable but two of the component years are reportable, and one of these years is flagged '+' and the other '=', then the core metric flag will be imputed as '+'; this suggests that greater value is placed on significant/material results over neutral ones, even though a null statistical inference is no less an inference than a non-null one. Secondly, it is not clear what is to happen when there are conflicting positive and negative flags for the component years. For example, if the core metric is non-reportable but two of the component years are reportable, and one of these years is flagged '+' and the other '-', then is the core metric flag to be imputed positively, negatively, neutrally, or not at all? This situation has not yet arisen in practice but, if and when it does, there does not appear to be a strategy in place to deal with it. Thirdly, this imputation approach was uncovered through discussion with DfE and OfS, and does not appear to be mentioned in the published TEF documentation. We would expect this methodological information to be made available to users.

We consider that the method of imputing missing flags for core metrics from their individual component years should be reviewed, and further thought be given to discontinuing this practice, and make a formal recommendation on this within Recommendation 17.

After the imputation methodology described above is carried out, metrics that are still non-reportable do not contribute to each provider's weighted sum of positive or negative flags; in effect, they are treated as though they carry neutral '=' flags. This implicit assumption of neutrality is an arbitrary one, and represents a crude form of imputation. If a provider has a relatively high number of non-reportable metrics then it is more difficult for it to achieve a Gold or Bronze categorisation at Step 1a, and it is more likely to remain in the Silver category compared to providers with fewer non-reportable metrics, all else being equal.

The intuitive reasoning outlined above is, at least to some extent, supported by the observed TEF data (see Table 10). For each of TEF Years Three and Four, we estimated a multinomial logistic model by regressing each provider's derived Step 1a starting point (using Silver as the reference category) on their number of reportable metrics (zero to six), and found a positive association between the number of reportable metrics and the odds of achieving a Step 1a starting point of Gold or Bronze rather than Silver. In TEF Year Three, the odds of achieving Gold rather than Silver increased by a factor of 2.2 for every additional reportable metric, while the odds of achieving Bronze rather than Silver increased by a factor of 5.3. In TEF Year Four, the odds of achieving Gold rather than Silver increased by a factor of 3.7 for every additional reportable metric, while the odds of achieving Bronze rather than Silver increased by a factor of 2.1. The odds ratios were significant at the 5% level for TEF Year Four but not for TEF Year Three, possibly reflecting the smaller number of providers in the sample (only providers that were assessed for a TEF award are present in the Year Three dataset, whereas the Year Four dataset includes all providers in the scope of TEF in England, irrespective of whether they were assessed, plus those from the devolved administrations that chose to participate).

We analysed the subject-level 2018/19 TEF pilot data in a manner largely analogous to that for the provider-level data as outlined above, except that we included random intercepts to account for intra-provider and -subject correlation (which necessitated the use of binary rather than multinomial logistic regression). As with the provider-level data, we found positive and statistically significant associations between the odds of achieving Gold or Bronze at Step 1a and the number of reportable metrics at the subject level (Table 11), although the estimated odds ratios were somewhat smaller in magnitude.

The analyses summarised in Tables 10 and 11 are intended to be descriptive in nature and are restricted to estimating only bivariate associations between the Step 1a starting point and the number of reportable metrics; our intention is not to construct a model that explains or predicts the Step 1a starting point in some broader sense, so a full set of model diagnostics is not reported here. In particular, it should be noted that we have not controlled for factors that could be confounded with the number of metrics reported by providers (such as provider size or type); identifying such factors is beyond the scope of this evaluation.

Table 10. Output of multinomial logistic regression associating Step 1a starting point with number of reportable metrics, provider-level TEF Years Three and Four

TEF Year	Step 1a starting point	Odds ratio (No. reportable metrics)	P-value
TEF Year Three	Gold	2.1608	0.3266
	Silver	-	-
	Bronze	5.3455	0.0629
TEF Year Four	Gold	3.6842	0.0337
	Silver	-	-
	Bronze	2.1159	0.0003

Note: models estimated on 86 observations for TEF Year Three and 396 observations for TEF Year Four

Table 11. Output of binary logistic regression associating Step 1a starting point with number of reportable metrics, subject-level 2018/19 TEF pilot

Step 1a starting point	Odds ratio (No. reportable metrics)	P-value
Gold	1.3587	<0.0001
Silver	-	-
Bronze	1.3811	<0.0001

Note: odds ratio for Gold estimated on 1,994 provider-subject combinations with a Gold or Silver starting point at Step 1a; odds ratio for Bronze estimated on 2,274 provider-subject combinations with a Bronze or Silver starting point at Step 1a

In light of the considerations above, we recommend that the overall approach to dealing with non-reportable metrics is reviewed and, at the very least, users should be made aware of the methodology, including its potential drawbacks, in a full and transparent way.

Recommendation 17: Review the approaches used for dealing with

- non-reportable metrics
- the imputation of missing flags from individual component years (including consideration of discontinuing this practice)

making the TEF documentation of these methods and approaches fully transparent.

3.9 Majority mode

Metrics are reported for each mode, that is separately for students studying full-time and those studying part-time. However, it is only the majority mode – that mode for which the headcount is higher – whose metrics are used in the Step 1a calculation, though providers (or subjects) with 'similar numbers of students' by mode are subject to additional consideration by the TEF Panel (OfS (2018.44a, paragraphs 22 to 25, and OfS (2018.44, paragraphs 264-265)).

It might be useful to devise a formulaic approach that combines the metrics – whether separately or overall for Step 1a – for full-time and part-time study in some suitably weighted way. At present, the experience of students in each provider's minority mode (for example, students undertaking part-time study at a university where most students are full-time, as is the case for most providers) is not captured in the Step 1a outputs. This idea is expanded upon in Section 4.3.3.

Recommendation 18: Consider a formulaic approach to combining the metrics for both full-time and part-time students.

3.10 Comments on the principle of combining metrics

The current TEF methods see six or nine metrics combined into a single rating for the provider or subject-within-provider. The process includes a formulaic approach in Step 1a, via weights and flags, followed by subjective appraisal and consideration by the TEF assessors and Panel. Of course, the contributing metrics' data are also available to the assessors, but currently none are categorised in the same way (Gold, Silver or Bronze) as the current overall measure.

Most of our considerations in this evaluation relate to there being a single, overall measure, but we briefly consider other options here. Possibilities include:

- not providing a rating at the overall level; rather, each contributing metric would need to be assessed and considered in its own right
- rating individual metrics as Gold, Silver and Bronze, or according to some other categorisation or rating scheme.

Further options exist about the level of aggregation of metrics:

- some metrics, notably those based on NSS data, are already an aggregation of responses to a number of individual questions on the survey
- various intermediate levels of aggregation could be produced, the most obvious choice being the three TEF Aspects of Quality (teaching quality, learning environment, and student outcome and learning gain).

However, other groupings are possible. For example, through judicious choice of input weights in a personal TEF calculator (as proposed in Section 3.5) a user could choose to equally weight Metrics 1, 3, and 5 and zero-weight Metrics 2, 4 and 6 and thus, in effect, create a new aggregate group. Indicators, benchmarks, differences, z-scores and flags could all be calculated for such aggregations.

In a purely mechanical sense, combining the TEF metrics is relatively simple. At present, this is achieved via flags, weights and rules for a benchmarked measure, and Section 4.2 considers an approach for a non-benchmarked measure. The combining of TEF metrics relies on having a set of pre-determined weights. Those need to be chosen and agreed subjectively (though that should be done by experts) as there is no formulaic approach for determining how 'Continuation' should be

weighed against 'The teaching on my course', for example. By contrast, official statistics times series used as indicators of the economy are routinely aggregated using weights, but these can be determined by conceptual requirements (actual patterns of expenditure can be used to determine weights for price indices, for example).

Naturally, there is an appeal in having a single, combined measure for a provider (or subject). It should assist assessors to make more consistent assessments, although the extent to which that works depends on the quality of the inputs to the measure, and how it is then considered alongside other information. Conversely, a combined measure has the potential to hide, or at least draw attention away from any variation that exist across metrics that should be considered separately. Even if a single, overall assessment is ultimately required, it may be better for assessors to consider the metrics separately to get an accurate and well-rounded overall view.

That the TEF metrics attempt to capture very differing concepts also does not help the case for the production and use of a single, combined measure. A relatively recent example of official statistics in which a single measure was **not** developed is personal well-being (harmonised questions on which are asked on the GO survey). Measures of personal well-being are reported separately for four topics – Happiness , Worthwhile, Life satisfaction, and Anxiety – and the reason for not producing a combined (composite) measure is the diversity of aspects they measure, and the need to consider them together to get a balanced view:

> "Office for National Statistics uses four questions to measure personal well-being and does not produce one composite measure. The questions were designed to measure distinct aspects of personal well-being (evaluative, eudemonic and affective). It is therefore not appropriate to combine these questions as they are all individually important and together they give a balanced approach to well-being." Source: ONS

Arguably, the TEF metrics could be viewed in a similar way.

It is difficult for us to give a statistical steer on the use of a combined measure in TEF, but think it might be useful for the issue to be considered further, and this should be done in the context of other recommendations we make about the Step 1a process. DfE has informed us that various research (for example, BIS (2016c, pages 33-34), DfE (2016b, pages 28-31) and DfE(2017c, chapter 8)) has been undertaken into TEF-award approaches and the usefulness of single, combined indicator. Naturally, the findings of these should be included in any further consideration.

Recommendation 19: In the context of the different core metrics capturing a diverse range of information, consider the usefulness of a single, combined measure in Step 1a, alongside the other recommendations we make about the Step 1a process.

4. Possible adaptations to Step 1a processes

This section considers various ways in which the Teaching Excellence and Student Outcomes Framework (TEF) Step 1a calculations could be amended, but largely staying with the current approach. Section 4.1 considers the creation of a net total value, comparing that with the current starting-point diagram and rules, and suggests how such a measure could be used to suggest starting points less rigidly than the current approach. Section 4.2 considers an analogous non-benchmarked measure, possibly to be presented alongside one that is benchmarked, and Section 4.3 presents an analysis of sensitivity of the current rules to variations in metric weights, flagging rules, and the consideration of both, rather than usually just one of the full- and part-time mode data.

4.1 Starting points under the current process

The 2017 and 2018 methods use the diagrams in Figures 1 and 2 (or equivalently, the rules behind them) to determine the starting points at Step 1a. A starting-point calculation has two inputs: the total value of positively flagged metrics, and the total value of negatively flagged metrics. In this section we consider a further calculation involving these totals, and present it from the perspective of the current 2017 and 2018 methods; however, it would also be applicable if those methods were adapted to include the suggestions made in Section 3 about relative weights and use of *prop* (proportion) functions.

4.1.1 A net total value: differencing the positive and negative flag values

As reviewers, our instinct was to want to combine these two inputs into a single value by simply deducting the total value of the negatively flagged metrics from the total value of the positively flagged metrics. Both the positive and negative totals are measures of the same thing, and so this approach seems right in principle. We shall refer to the resulting quantity as the **net total value** (of flagged metrics). This quantity can be shown on the starting-point diagrams.

Figure 5 shows the net total values for both the 2017 and 2018 methods; these have replaced the ‘G’, ‘G/S’, ‘S’, ‘S/B’ and ‘B’ labels in Figures 1 and 2, though the shading of the different starting point categories has been retained. (Note that the recommended relative metric weights are not now shown). All starting-point cells on the same top-left to bottom-right (‘\’) diagonals share the same net total value, with more positive scores to the right or above, and more negative scores to left or below.

The derived net total value, though seemingly natural and intuitive, produces anomalies and inconsistencies with the current methods. This should not come as a surprise as the lines of constant net total value are diagonal, whereas the starting-point boundary thresholds are horizontal in the 2017 method, and much closer to horizontal than on the diagonal in the 2018 method.

In particular, it is worth highlighting the amount of overlap in net total value that is present between starting-point categories; this can be seen in Figure 6. Under the 2017 method, the starting point category ranges are:

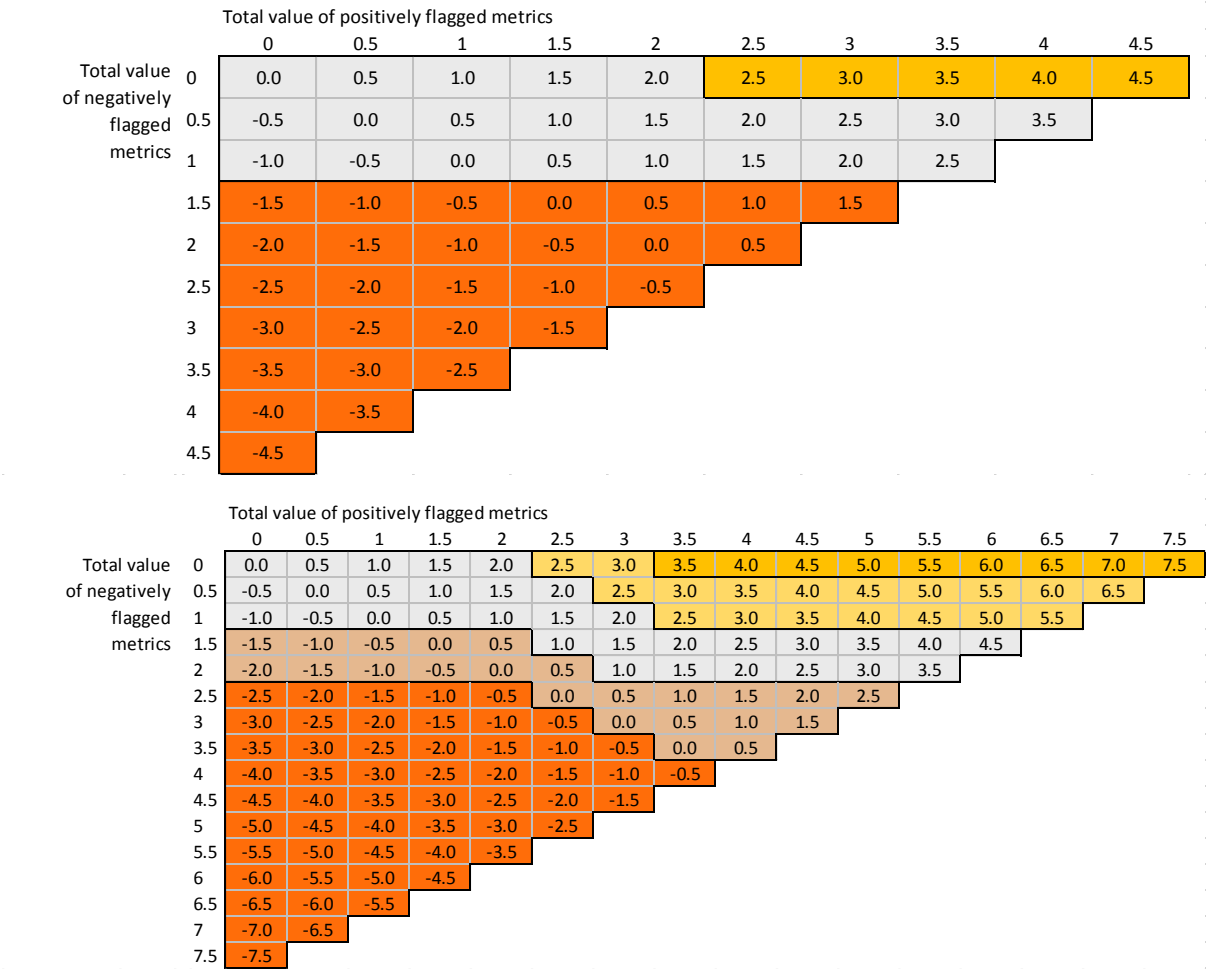
Bronze –4.5 to +1.5 | Silver: –1.0 to +3.5 | Gold: +2.5 to +4.5

Under the 2018 method, the starting point category ranges are:

Bronze: –7.5 to –0.5 | Silver/Bronze: –2.0 to +2.5 | Silver: –1.0 to +4.5 |

Gold/Silver: +2.5 to +6.5 | Gold: +3.5 to +7.5

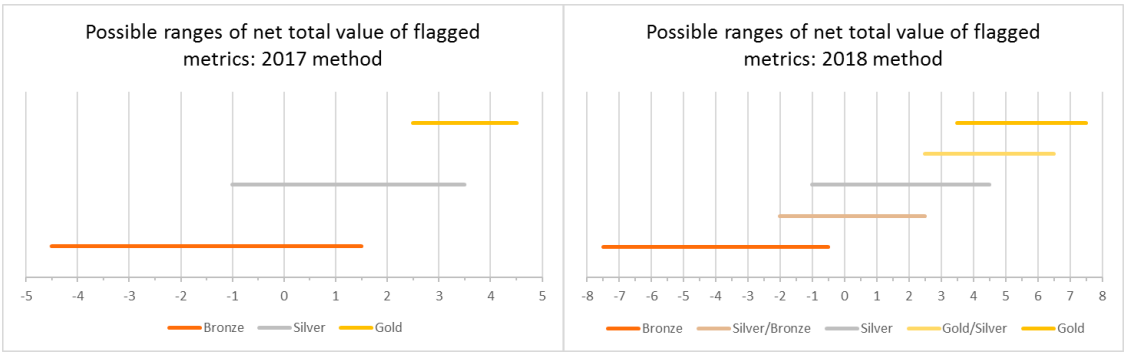
Figure 5. 2017 method and 2018 method starting-point diagram showing net total value of flagged metrics



Total value of negatively flagged metrics \ Total value of positively flagged metrics	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5
0	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
0.5	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	
1	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5		
1.5	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5			
2	-2.0	-1.5	-1.0	-0.5	0.0	0.5				
2.5	-2.5	-2.0	-1.5	-1.0	-0.5					
3	-3.0	-2.5	-2.0	-1.5						
3.5	-3.5	-3.0	-2.5							
4	-4.0	-3.5								
4.5	-4.5									

Total value of negatively flagged metrics \ Total value of positively 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	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5
0.5	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	
1	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5		
1.5	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5			
2	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5				
2.5	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5					
3	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5						
3.5	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5							
4	-4.0	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5								
4.5	-4.5	-4.0	-3.5	-3.0	-2.5	-2.0	-1.5									
5	-5.0	-4.5	-4.0	-3.5	-3.0	-2.5										
5.5	-5.5	-5.0	-4.5	-4.0	-3.5											
6	-6.0	-5.5	-5.0	-4.5												
6.5	-6.5	-6.0	-5.5													
7	-7.0	-6.5														
7.5	-7.5															

Figure 6. Starting-point category ranges of net total value



We note that the boundary values of all starting-point categories can, in theory, be realised (that is, there exist combinations of the metric weights which, if the metrics are flagged positively or negatively, could result in the combination of the boundary net total value and starting-point category shown).

We also draw attention to the fact that, in the 2018 method, it is not just the borderline categories (Gold/Silver and Silver/Bronze) that overlap with the main categories (Gold, Silver, Bronze), but the main categories have some overlap with each other. For example, a net total value of 3.5 can be sufficient for a starting point of Gold, whereas a (higher) net total value of 4.5 may result in a starting point only of Silver. Examples of this sort, assuming a net total value is desirable, appear wrong in principle to the reviewers and we wonder whether this could leave TEF open to appeals.

4.1.2 Penalties for negative flags

We have seen that defining a single measure as the net difference between the positive and negative flag values produces results inconsistent with the current starting-point diagrams. Therefore, we now consider whether there are alternative single, 'net' measures that are more consistent, and would therefore provide an alternative interpretation of the current rules. What we consider in this section should be seen as just that – an interpretation of the current rules – rather than any statement on whether it is any more, or any less, the 'right thing to do'.

The TEF documentation seems to regard negative flags as something particularly important to consider, and as something very undesirable. A requirement for a Gold stating position under both the 2017 and 2018 methods is that there are no negative flags. That is to say, that no matter how many positive flags of any value are present, just one negative flag of any value renders a Gold Step 1a starting point unattainable. The Gold/Silver borderline category introduced in the 2018 method allows up to 1.0 in negative flag values.

We find a similar bearing of the negative scores placed on the definitions for Bronze: its starting-point region under the 2017 method is reached if the total of the negative flag values exceeds 1.5 'regardless of the number of positive flags'. However, in the 2018 method, the requirement is relaxed a little, as seen in the diagram (Figure 6) where the previously horizontal upper threshold for Bronze is replaced with a horizontal section followed by a diagonal decline when in the presence of greater positive totals.

There is no simple, mathematical formula that captures neatly all the intricacies of the starting-point diagrams and their rules, but a simple amendment to the 'net' approach, would be to vary the relative contributions of the positive and negative flags.

We could redefine the net total value as the positive total less some fixed multiple of the negative total:

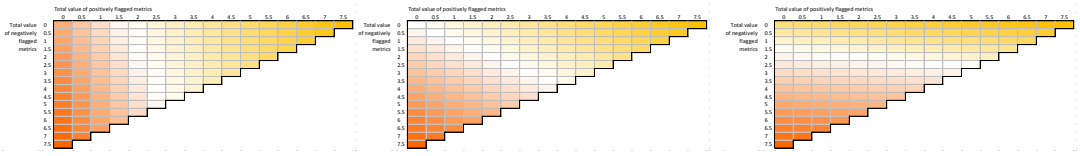
$$\text{net total value'} = \text{positive total} - k \times \text{negative total},$$

where k is some chosen constant. Setting $k = 1$ would give equal emphasis on positive and negative flags (as we have investigated already), whereas setting $k > 1$ would penalise the negative metrics more harshly than the positive metrics reward. (For completeness, setting $k < 1$ would penalise the negative metrics less harshly). We don't investigate this formula further for the 2017 method, as the category boundaries are largely horizontal. In such a formula, the positive total would be ignored (similar to making k very large); instead we consider just the 2018 method diagram.

Three indicative diagrams (Figure 7) follow for the 2018 method with $k = 0.2$, $k = 1$ and $k = 5$. For example, a value of $k = 5$ is equivalent to stating that a positive flag on the 'Continuation' metric has (positive) value 2.0, whereas a negative flag has a (negative) value of 10.0. The redefined net total values are of now of no interest in themselves (they are arbitrary and have new possible totals that are dependent upon the value of k; therefore, they are omitted from the diagrams). In these

diagrams, we have used an automated grading tool to differentiate the relative net total values, with graduated shading from bronze through silver to gold; note there are no longer any 'regions' defined with specific boundaries, as these would need to be re-drawn anyway to accommodate different possible totals. In these diagrams, it is the patterns of shading that are of interest, and these should be compared with those in the current 2018 method starting point diagram (Figure 2).

Figure 7. Graduated scaling of the redefined net total value



Graduated scaling of the re-defined net total value, in which different contributions from the positive and negative flags are allowed: net total value' = positive total – k x negative total. From left-to-right, k = 0.2 (positive scores are more dominant), k = 1.0 (corresponding with equal contribution, as investigated in Section 4.1.1), and k = 5.0 (harsher penalisation of negative scores).

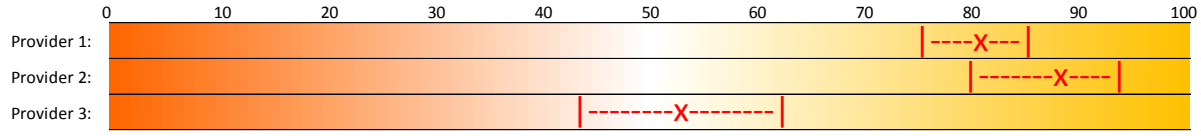
The middle diagram of Figure 7 (equal emphasis of positive and negative scores) illustrates the diagonals of equal net total value already discussed. The patterns illustrated clearly differ from those seen in the current 2018 and 2017 methods' starting-point diagrams. However, the pattern shown in the third diagram is visually close to the current 2018 method rules-based starting points (gold along the top edge, strongest on the right; bronze in the bottom corner, with an upper edge that declines towards the right; and silver in between). This suggests that a simpler and arguably more transparent single measure, based on the net total value, can be found that would deliver similar results to the current rules. Whether that is desirable, is another question. Alternatively, the value of *k* could be selected based on expert judgement, possibly on non-empirical grounds, as long as the process by which *k* is selected is made sufficiently transparent to users.

4.1.3 Further thoughts on net measures

Perhaps the biggest appeal of the net total value measure is that, assuming its implementation is combined with the replacement of binary flag scores by mathematically-neat *prop* functions, is the scope it then allows for approximating standard errors for that combined indicator, although we consider that would not be a straightforward task. Having a measure of statistical uncertainty presented with the Step 1a starting-point outcome would be extremely useful. If there is an appetite for some net-value, overall measure, it would clearly need to be developed further and tested thoroughly. It could also take into account weighted contributions from both full-time and part-time modes.

For presentation of the Step 1a starting point, the final score may be linearly rescaled again, for example to put it on a 0 to 100 range, with the confidence interval (assuming one can be approximated) shown. Under these conditions, there may no longer be a need to define starting-point categories with specified boundaries, but a colour-graded scale could be used instead. Figure 8 provides an illustrative example. TEF assessors could then work from that location (and confidence interval) to form their initial hypothesis.

Figure 8. Illustrative example of a graded-scale Step 1a output with confidence intervals



Use of a scale, rather than categorising the Step 1a outcome, might capture better some essence of uncertainty in the Step 1a process. How workable such a proposal would be would require testing with TEF assessors. Most notably, the requirement for TEF assessors and Panel members to digest a large volume of empirical data raises the question of whether visual flags should be retained alongside a continuous net measure to draw attention to cases of notably strong or weak performance; both approaches in combination are possible.

A wider, but related issue is whether the final TEF output should be a Gold/Silver/Bronze categorisation at all. Would it be more useful, and transparent, for final users of TEF outputs (potential students and providers, for example) to work with those same numbers available to the TEF assessors and Panel? That is, separate and combined metric values could be presented instead of a three-category output. We note that the Royal Statistical Society (RSS) comments similarly (RSS, 2018, Section F). If the separate assessment of different types of provider were ever to be implemented (see Recommendation 34 in the Conclusions (Section 11)), re-considering the definitions or labelling of Gold, Silver and Bronze may be desirable anyway.

We recommend that serious consideration is given to the further development of the rules and processes used to derive the Step 1a starting points, as both the 2017 and 2018 methods have the capacity to produce unintuitive or inconsistent ratings. One approach for achieving this would be the development of a net-value measure.

Recommendation 20: The Step 1a methods should be developed further, and in a holistic way, noting the other recommendations made on specific aspects, and that development should include the consideration of a net total value measure, proportion functions and approximation of confidence intervals, if possible.

4.2 Development of a single, overall measure that does not compare against benchmarks

The current Step 1a approach, especially if taken further to produce an overall net measure in any of the ways suggested in Section 4.1, can be considered as providing an assessment of a provider's performance in comparison to its benchmarks. It is, therefore, a relative measure. Some would see this as a positive feature – it 'levels the playing field' – but others may criticise it for not providing an absolute indicator, one for which an improvement in quality by the provider should result in an improved indicator, and not be affected by changes in other providers' performance.

Thus, we might also consider the values of the indicators themselves, rather than the differences between them and their benchmarks as important. TEF pays some attention to this already in terms of high and low 'absolute' values (discussed further in Section 5.1), but considers the metrics individually; there is no overall, combined 'absolute' measure in the way there is for the benchmarked scores in Step 1a.

It may be useful for users to be able to consider both a single, combined indicator value produced for each provider (and subject) that is not benchmarked, as well some net value based on comparisons with benchmarks. Having both non-benchmarked and benchmarked measures side-by-side is likely to show clearly what effect benchmarking has, and should reinforce that the current measure is, indeed, benchmarked. Pairs of non-benchmarked and benchmarked Step 1a outputs could be presented together on a diagram too, for example expanding on that already shown in Figure 8.

A non-benchmarked measure could be produced easily, given that each indicator is currently a proportion and therefore measured on the same scale, and each indicator is a measure of something positive or desirable:

$$p_{\text{combined}} = \frac{\sum_m w_m p_m}{\sum_m w_m}$$

where the sum is over all six (2017 method) or nine (2018 method) core metrics, *m*. Given it would be based on data (the metrics' indicators) already publicly available in the TEF workbooks, producing this non-benchmarked measure would be a simple task, and one better done by the statistical producer than by individuals or other organisations as it then provides an authoritative measure, and reduces the risk of user-error in the calculations. How such a measure should then be used – whether it should feature in the Step 1a calculations, what weight should be placed upon it, and so on, by TEF Panel members, potential students and providers – should be considered further and may benefit from the production of further guidance.

It would be desirable to produce a confidence interval around this measure too, which should be an easier task than for the benchmarked net total value, but still not straightforward. For both benchmarked and non-benchmarked measures, the metrics are correlated (as shown in Section 3.5) and a number of the metrics are derived from the same sample and data source, and are thus not independent. Covariances between the metrics, a necessary input for deriving confidence intervals for a combined measure, would need to be approximated and may need to be estimated empirically; we have not investigated this proposal further. In addition, considerations of multiple-comparison tests would be relevant here too in deciding the appropriate width of any confidence intervals.

A non-benchmarked measure could also be included in any 'personal TEF calculator', as suggested in Section 3.6.

Recommendation 21: Consider developing an analogous, non-benchmarked version of a combined indicator, which could be presented alongside a benchmarked version in Step 1a; further guidance on its use may also be required.

4.3 Sensitivity analysis of alternative options in Step 1a

We have conducted some analysis using the published TEF Year Three metrics data to assess the sensitivity of the Step 1a starting points to changes in the process, using some of the approaches suggested earlier in this section. In doing so, we have applied the published 2017 methodology to the published metrics data to produce our own Step 1a starting points. It should be noted that our derived TEF Year Three starting points do not perfectly match the confirmed starting points supplied to us by the Office for Students (OfS); we derived 19 Bronze, 62 Silver and 5 Gold starting points, compared with confirmed starting points of 22 Bronze, 58 Silver and 6 Gold. We believe this is because of differences between the published methodology and the way in which the process is implemented in practice (for example, imputation of flags for non-reportable core metrics using individual component years, as discussed in Section 3.8).

4.3.1 Changes to weights

We assessed the sensitivity of the Step 1a starting points to the choice of metric weights, whereby the current weighting scheme (0.5 for each of the three National Student Survey (NSS) metrics, and 1.0 for each of the other three) was amended in turn to each of the following schemes (weights were scaled to sum to 4.5 in all cases):

- equal weights
- weights proportional to sample size (pooled across providers)
- weights inversely proportional to variance (pooled across providers)
- all weight on Metric 1 ('The teaching on my course'), no weight on the rest
- all weight on Metric 2 ('Assessment and feedback'), no weight on the rest
- …
- all weight on Metric 6 ('Highly-skilled employment or further study'), no weight on the rest.

The resulting weighting schemes and derived Step 1a starting points are summarised in Table 12. Applying equal weights generally resulted in a similar frequency distribution of starting points to the current approach, while weighting in proportion to sample size or inversely in proportion to variance resulted in fewer Gold and Bronze outcomes. The impact on the frequency distribution of putting all weight on just one metric was variable, depending on the metric chosen, but generally there were more Gold and fewer Bronze outcomes. These results should be treated with caution, as they tend to mask a significant number of changes to outcomes for individual providers. For example, if all weight is placed on 'Highly skilled employment or further study', the Step 1a starting point changes for 29 out of the 86 assessed providers; indeed, the starting point changes for a substantial proportion of the providers for all tested scenarios other than that of equal weighting. These results suggest that the TEF Year Three Step 1a starting points are somewhat sensitive to the choice of weighting scheme. It should be noted that we are not suggesting that any of the assessed weighting schemes are in any sense optimal. We reiterate our earlier recommendation from Section 3: the choice of weights should be decided by an expert panel, and the approach to selection should be transparent.

Table 12. Derived Step 1a starting points under alternative weighting schemes, TEF Year Three

Metric weight	Metric weight						Step 1a starting point (no. of providers)			No. of starting points changed
	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5	Metric 6	Gold	Silver	Bronze	
Current weight (2017 method)	0.50	0.50	0.50	1.00	1.00	1.00	5	62	19	-
Equal weights	0.75	0.75	0.75	0.75	0.75	0.75	4	62	20	2
Proportional to sample size	0.71	0.71	0.71	1.08	0.64	0.64	4	72	10	10
Inversely proportional to variance	0.61	0.37	0.48	1.55	1.14	0.35	4	68	14	20
All on Metric 1: 'The teaching on my course'	4.50	-	-	-	-	-	5	74	7	18
All on Metric 2: 'Assessment and feedback'	-	4.50	-	-	-	-	8	67	11	21
All on Metric 3: 'Academic support'	-	-	4.50	-	-	-	11	63	12	21
All on Metric 4: 'Continuation'	-	-	-	4.50	-	-	8	65	13	25
All on Metric 5: 'Employment or further study'	-	-	-	-	4.50	-	4	73	9	17
All on Metric 6: 'Highly skilled employment or further study'	-	-	-	-	-	4.50	14	47	25	29

4.3.2 Changes to thresholds used in flagging rules

We assessed the sensitivity of the Step 1a starting points to the choice of thresholds used to derive the '+'/'++'/'-'/'--' flags, whereby the current ruleset ($|d|>2$ and $|z|>1.96$) was replaced with various combinations of $|d|$ (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0) and $|z|$ (1.645, 1.960, 2.576); the chosen z-scores correspond to the 10%, 5% and 1% two-tailed significance levels, respectively.

As summarised in Table 13a, the number of providers with a Gold starting point at Step 1a ranged from 2 to 8 (compared with 5 under the current ruleset), whilst those with a Bronze starting point ranged from 9 to 26 (compared with 19 under the current ruleset). Of the 86 providers assessed in TEF Year Three, the total number whose Step 1a starting point changed was maximised at 14 (when the threshold for $|d|$ was increased from 2.0 to 3.5 or 4.0, and that for $|z|$ was increased from 1.960 to 2.576), as summarised in Table 13b. Of course, if the thresholds for $|d|$ and $|z|$ were moved even further from their existing values of 2.0 and 1.96, we would expect the number of Step 1a starting point changes to be even more pronounced. As with the choice of weighting scheme, these results suggest that the TEF Year Three Step 1a starting points are somewhat sensitive to the choice of thresholds employed in the flagging rules.

Table 13a. Derived Step 1a starting points (Bronze / Silver / Gold) under alternative *d* and *z* thresholds, TEF Year Three

Absolute *d* threshold	Absolute z-score threshold		
	1.645	**1.960**	**2.576**
1.0	26 / 52 / 8	24 / 55 / 7	18 / 62 / 6
1.5	23 / 57 / 6	21 / 59 / 6	16 / 65 / 5
2.0	21 / 60 / 5	**19 / 62 / 5**	15 / 67 / 4
2.5	17 / 65 / 4	15 / 68 / 3	11 / 72 / 3
3.0	16 / 67 / 3	14 / 69 / 3	10 / 73 / 3
3.5	16 / 68 / 3	13 / 70 / 3	9 / 74 / 3
4.0	15 / 59 / 2	13 / 71 / 2	9 / 74 / 3

Table 13b. Frequency of Step 1a starting points that changed between the current ruleset and alternative *d* and *z* thresholds, TEF Year Three

Absolute *d* threshold	Absolute z-score threshold		
	1.646	**1.960**	**2.576**
1.0	10	7	12
1.5	5	3	9
2.0	2	-	7
2.5	7	6	12
3.0	9	7	13
3.5	10	8	14
4.0	11	9	14

4.3.3 Weighting together metrics for full-time and part-time students

In the current TEF process, only the metrics for each provider's majority mode of study are considered when deriving the Step 1a starting points. This means that the experiences of students in the minority mode (part-time for most providers) are not taken into account in Step 1a. We examined the impact on the Step 1a starting points of combining each provider's metrics for full- and part-time students, by weighting together the mode-of-study splits for *d* and std(*d*) in proportion to the number of full- and part-time students in the denominator.

Under the current approach, the distribution of Step 1a outcomes amongst the 86 providers assessed in TEF Year Three consisted of 19 Bronze, 62 Silver and 5 Gold starting points. After weighting together the metrics for full- and part-time students, this distribution changed to 22 Bronze, 59 Silver and 5 Gold starting points. A total of 9 providers experienced a change in starting point (three promoted and six demoted). These results suggest that weighting together the metrics for full- and part-time students would have a small, though non-negligible, effect on the Step 1a starting points compared with the current approach of using solely the majority mode of study.

5. Contextual data

In this section we consider the contextual data, that is the data that do not directly inform the formulaic calculation of the Step 1a starting point, but that are considered by Teaching Excellence and Student Outcomes Framework (TEF) assessors and the TEF Panel in deriving and agreeing the final assessment(s).

We have examined some aspects of such data in more detail, especially where there are processes or assumptions made that differ from those of the core metrics. A more general consideration of the role the contextual data play in forming the final assessments in given in Section 10.1, in which we report on our findings of interviews with TEF assessor and Panel members.

5.1 High and low absolute values

Special symbols are used to denote large and small values of an indicator, *p*, of a metric. The symbols used – and showing in the TEF workbooks – are '*' and '!' respectively; the main TEF documents describe such cases as very high and very low 'absolute' values.

The 2017 method for defining high and low values (DfE, 2017a, paragraphs 5.65 to 5.66) seems fairly straightforward to understand – simply, all those providers (and we understand that to include those providers not being assessed for TEF at the time), in the top or bottom deciles of the given mode (full-time or part-time), subject to meeting a minimum criterion on the number of students, receive the '*' or '!' symbols.

The 2018 method (OfS, 2018.44, paragraphs 250 to 252; OfS, 2018.44a, paragraphs 87 to 96) seems to be a development of the 2017 method and is more complicated. In addition, colours are added to the symbols: '*' may be green or grey, and '!' may be blue or grey. We understand that at provider level, the process for defining the high values is as follows (low values are defined analogously, and we omit that description):

> For any given metric, the providers are ranked according to the value of the indicator, *p*.
>
> All those providers in the top decile will receive the symbol '*', but the colour is not yet defined.
>
> The value of the indicator of the lowest-ranked provider in the top decile is used as a threshold. Then:
>
> a provider whose 95% confidence interval for the indicator sits entirely above the threshold, or has *p* = 100% and more than 100 students), receives a **green** '*' symbol, subject to there being no contradictory benchmarked flags in the core metric or splits.
>
> a provider not meeting the criteria above (that is, its confidence interval overlaps the threshold, or *p* = 100% but the number of students is 100 or fewer, or there are contradictory flags) has its '*' symbol coloured **grey**.

Although we didn't find it clearly described in the documentation, we understand that, for subjects, the subject-level indicator data are compared against thresholds set at provider-level (rather than setting subject-specific thresholds) for each metric.

We have a number of comments on this approach (mainly the 2018 method), which we outline in the following subsections, and make an overall recommendation at the end

5.1.1 Principles

There is clearly some merit in trying to identify the highest and lowest providers in terms of the values of their (non-benchmarked) indicators. In this sense, the aims of this process are reasonable. We commented on use of non-benchmarked measures in Section 4.2.

5.1.2 Documentation

Even after much study of the documentation, we are not convinced we have fully and correctly understood all the criteria and fairly complicated rules. Also, the documentation does not state what definition of confidence interval for *p* is used, though the Office for Students (OfS) has confirmed to us it is a Wilson-score interval.

We are not quite sure how the comment in the documentation 'for any population size being referred to' (OfS, 2018.44a, paragraphs 90a and 93a) should be interpreted. It may be in reference to a fixed and finite student population, or some super-population if taking the approach used elsewhere; it might be suggesting that all student population or sample sizes (whether up to or above 100) should be considered; or it might mean 'for any provider or subject'.

Use of the term 'absolute' (in absolute high or low values) is possibly incorrect or misleading: this method can only identify the 10% of the best/worst providers in comparison to others according to a set of rules. In that sense, the values identified as high or low are not absolute, but relative.

5.1.3 Methods and process

Our comments in this area are as follows:

Comment 1. A smaller provider (in terms of student numbers) that is in the top or bottom decile, is less likely than a large provider in that decile to be given a green '*' symbol (called a star in DfE (2017a) and OfS (2018.44a), and an asterisk in OfS (2018.44)) or a blue '!' symbol (exclamation mark) rather than a grey symbol, even if their non-benchmarked metrics data are the same.

This is simply because confidence intervals based on smaller samples are wider, and therefore more likely to cross any given threshold. We note that if the actual, finite population size were used – instead of assuming a super-population – the width of the confidence interval would tend to zero as the responding sample size approached the population size.

As an illustrative example, it is possible to find in the available National Student Survey (NSS) data an example of full response from a population of 11 students, all of whom responded positively to the question asked (thus $p = 100\%$). In a paradigm of a fixed and finite population – the assumption common in many official statistics outputs – the confidence interval around this estimate would have zero width: the target population has been fully enumerated, and so there is no sampling variability. However, the approach adopted in TEF and by the NSS is to assume these 11 students are a realisation from a much larger population; there is therefore uncertainty present in the estimate, and that results in the quoted confidence interval around the estimate of $p = 100\%$ being (44%, 100%). A wide interval such as this, would be very unlikely to receive a green '*' symbol despite every student in the target population responding and responding positively. This issue would be more prevalent in subject-level data (where sample sizes will be smaller) than for provider level, and would become more of an issue if the subject breakdown were ever to use the more detailed Common Aggregation Hierarchy (CAH) Level 3 classification than CAH Level 2.

Comment 2. That the value of the threshold is determined by a provider (as is always the case for determining deciles), rather than being a fixed number, could affect the colour of the symbol applied to other providers' indicators and could make comparisons over time more difficult.

The value of the threshold, given it is some provider's indicator value, is a point-estimate and therefore is subject to sampling- and other errors. It may not be an accurate representation of the (underlying or true) 90th percentile of the measure, and its confidence interval may be relatively wide. We note that the 2018 method necessitates that the calculation of the threshold value will be based only on providers with at least 100 students, but this minimum sample size could still leave a reasonable amount of uncertainty about the value of the threshold. Using the same (Wilson score with $z = 2.17$) calculations as used on NSS confidence intervals, we note that with a sample size of $n = 100$:

- a threshold estimate of $p = 50\%$ would have a confidence interval of (40%, 60%)
- a threshold estimate of $p = 75\%$ would have a confidence interval of (65%, 83%)
- a threshold estimate of $p = 90\%$ would have a confidence interval of (82%, 95%)

At overall (that is, across all subjects) level, this is unlikely to represent a real problem, as there would be relatively many other providers, and the boundary-provider's nearest-neighbours' indicators would likely give similar estimates of that 90th percentile. However, if the number of providers were smaller, there might be larger differences between the boundary-defining provider and its nearest neighbours' indicators. Thus, in principle, the actual (unknown) true indicator of one provider could have an effect on the high and low symbol markers given to other providers. Again, we see that it is particular cases of providers near thresholds that provide possible inconsistencies in the approach or otherwise undesirable results.

At subject-level, the risks of small counts of providers is mitigated by applying provider-level (that is, across all subjects) thresholds to the subject-level data, rather than recalculating those thresholds using only data and providers relating to the subject in question. However, this introduces the issue of whether the provider-level thresholds are really appropriate for all the subjects taught by the provider. There is clearly a trade-off between robustness (having flags that are not affected by small-sample problems) and relevance (having thresholds that are appropriate for specific subjects).

We note that options for deriving the thresholds for the subject-level flags have already been considered (DfE, 2018), with some justification given for applying the provider-level thresholds. This lay largely around achieving consistency between provider-and subject-level awards, and for the symbols ('*' and '!') to reflect absolute rather than relative performance. However, the approach adopted for the pilot leads to inconsistent methodologies being applied between the provider- and subject-levels, and there remains the more general issue of the approach being based on relative rather than absolute performance (see Section 5.1.2). We suggest there is merit in reconsidering this issue as part of a wider reflection on the subject-level pilot data and methods (see Recommendation 33).

Comment 3. We note (OfS, 2018.44a, paragraphs 92b and 95b; OfS, 2018.44, paragraphs 251b and 252b) that, under particular conditions, a green star [blue exclamation mark] should be considered in a similar way to a positive [negative] flag in determining the final position of the initial hypothesis. However, this information is not taken into account in the formulaic determination of the Step 1a starting point (and it could be); rather it is left for the TEF assessors and Panel to note this and take account of it in their application of expertise in determining the final outcome.

Our recommendation covering various aspects of 'high and low absolute values' discussed throughout Section 5.1 follows.

Recommendation 22: The documentation and descriptions of 'very high and very low absolute values' and their methods should be made clearer and more transparent. The appropriateness of using provider-level thresholds for each specific subject should also be reviewed.

5.2 Split metrics

Whilst the formulaic Step 1a starting point is driven entirely by the core metrics, which are the main focus of this evaluation, Step 1b is informed by split metrics (as well as other sources of empirical evidence, which are combined in a non-formulaic manner). The split metrics relate to student subgroups (domains), and disaggregate each provider's six core metrics into categories defined according to 10 variables. This section considers provider-level split metrics for TEF Years Three and Four, using the 2017 method; subject-level splits for the 2018/19 pilot using the 2018 method are considered in Section 5.3.

Disaggregating the core metrics by the splits reduces the sample size (the number of students in the denominator) for each provider-metric combination. In TEF Years Three and Four, whilst the mean sample sizes across all provider-metrics within each split category appeared reasonable, the smallest samples consisted of fewer than 30 students (the 'textbook' minimum sample size required for the Central Limit Theorem to apply) within all splits except 'Welsh medium'. The percentage of provider-metric combinations that had fewer than 30 students in their sample was at least 10% in 9 of 27 split categories in TEF Year Three and 15 of 27 split categories in TEF Year Four (see Table 14). This raises questions over the reliability of the statistical inferences for some of the split metrics, and consideration should be given to removing splits with a high prevalence of small sample sizes, or at least collapsing their categories.

Recommendation 23: Consideration should be given to removing splits with a high prevalence of small sample sizes, or at least collapsing their categories.

Splitting the metrics data into domains also has the effect of reducing the number of metrics that are reportable by each provider within each split. When considering just the core metrics, all providers in the published TEF Years Three and Four datasets reported at least one metric, and the majority reported all six metrics. When turning our attention to the split metrics (see Table 15), fewer than half of providers in TEF Year Three were able to report all six metrics in 19 of the 27 split categories, whilst more than half of providers reported no metrics at all in eight splits. Non-reportable metrics were less of a problem in TEF Year Four, with at least half of providers reporting all six metrics in 15 of the 27 split categories, and fewer than half of providers reporting no metrics in all but two splits.

Table 14. Split metrics where at least 10% of samples consist of fewer than 30 students, TEF Years Three and Four

TEF Year	Split category	Min. sample size	Mean sample size	% samples <30 students
TEF Year Three	Disabled: Yes	12	351	21.5
	Domicile: Other EU	18	217	13.6
	Domicile: Non-EU	10	453	16.7
	Ethnicity: Asian	10	474	19.7
	Ethnicity: BME	10	576	18.8
	Ethnicity: Other	10	230	16.7
	IMD: Disadvantaged	10	574	10.5
	POLAR: Disadvantaged	10	550	12.5
	POLAR: Not disadvantaged	10	1,207	10.8
TEF Year Four	Age: Mature	10	619	10.9
	Disabled: Yes	10	382	21.2
	Domicile: Non-EU	10	449	13.5
	Domicile: Other EU	10	243	20.0
	Ethnicity: Asian	10	553	21.9
	Ethnicity: BME	10	1,003	14.6
	Ethnicity: Black	10	354	18.6
	Ethnicity: Other	10	238	24.2
	IMD: Disadvantaged	10	908	10.7
	Level of study: First degree	10	2,749	10.5
	Level of study: PG-UG border	10	544	18.3
	POLAR: Disadvantaged	10	554	15.9
	POLAR: Not disadvantaged	10	1,387	13.2
	Year: 2	10	795	11.8
	Year: 3	10	824	12.9

Notes:

- *Each observational unit is a provider-metric pair*
- *BME: black and minority ethnic; EU: European Union; IMD: Index of Multiple Deprivation; PG-UG: postgraduate-undergraduate; POLAR: Participation of Local Areas*
- *Descriptions of the split categories can be found in the Teaching Excellence and Student Outcomes Framework Specification, October 2017 (DfE, 2017a)*

Table 15. Percentage of providers with 0 or 6 split metrics reportable, TEF Years Three and Four

Split category	TEF Year Three		TEF Year Four	
	0 metrics reportable	6 metrics reportable	0 metrics reportable	6 metrics reportable
Year: 1	12.3	47.9	4.7	62.3
Year: 2	2.7	46.6	7.4	62.9
Year: 3	2.3	60.5	5.6	62.7
Sex: Female	3.5	53.5	8.5	62.0
Sex: Male	7.1	55.3	10.8	52.3
Age: Mature	5.8	55.8	11.7	57.6
Age: Young	5.8	51.2	12.2	58.9
Disabled: Yes	23.5	36.5	21.5	58.9
Disabled: No	3.5	46.5	11.4	60.8
Domicile: UK	3.5	0.0	26.6	0.0
Domicile: Other EU	69.7	0.0	57.6	0.0
Domicile: Non-EU	54.8	0.0	53.2	0.0
Ethnicity: Asian	39.5	24.7	39.3	35.4
Ethnicity: BME	22.4	37.6	31.5	28.8
Ethnicity: Black	51.3	22.5	44.1	30.6
Ethnicity: White	3.5	46.5	17.3	44.5
Ethnicity: Other	53.0	22.9	42.3	30.9
IMD: Disadvantaged	7.0	54.7	14.9	44.1
IMD: Not disadvantaged	4.7	57.0	11.2	51.7
POLAR: Disadvantaged	29.4	38.8	21.4	55.0
POLAR: Not disadvantaged	8.2	48.2	14.8	59.3
Level of study: First degree	15.0	51.3	13.2	49.6
Level of study: Other undergraduate	9.9	46.9	17.6	35.5
Level of study: PG-UG border	52.5	27.5	27.3	32.1
Welsh medium: 0-5	70.0	30.0	0.0	100.0
Welsh medium: 5-40	70.0	20.0	0.0	100.0
Welsh medium: Over 40	70.0	20.0	0.0	100.0

Notes:

- *Each observational unit is a provider-metric pair*
- *BME: black and minority ethnic; EU: European Union; IMD: Index of Multiple Deprivation; PG-UG: postgraduate-undergraduate; POLAR: Participation of Local Areas*
- *Descriptions of the split categories can be found in the Teaching Excellence and Student Outcomes Framework Specification, October 2017 (DfE, 2017a)*

As stated in Section 3.3, the multi-dimensional nature of the benchmarking grid (the cross-tabulation of benchmarking factors and providers) raises the possibility of having a small number of providers in a cell and subsequent self-benchmarking (whereby a provider makes a large contribution to its own benchmark). Cases such as these make it more difficult for the provider to be flagged positively or negatively against a particular metric (Forster, no date). This issue is accentuated when the metrics data are split into domains, particularly for certain splits (see Table 16). The maximum contribution by a provider (for any of the six metrics) within a split exceeded 20% for 23 of the 27 split categories in TEF Year Three, and it exceeded 50% for six of them. Whilst the mean percentage contributions to provides' benchmarks were generally similar between TEF Years Three and Four for every split category, the maximum contributions were notably greater in TEF Year Four, where they exceeded 50% (and were generally well in excess of this) for all but one split category.

Table 16. Providers' contribution to their own benchmark (%), split metrics, TEF Years Three and Four

Split category	TEF Year Three		TEF Year Four	
	Mean	Maximum	Mean	Maximum
Year: 1	4.4	55.3	4.4	79.6
Year: 2	4.3	47.9	4.9	78.9
Year: 3	4.8	48.6	4.9	75.9
Sex: Female	4.0	36.8	4.5	79.2
Sex: Male	4.4	47.2	4.6	77.6
Age: Mature	4.3	38.8	4.8	82.2
Age: Young	3.1	27.5	3.1	73.9
Disabled: Yes	5.9	32.5	6.0	83.6
Disabled: No	3.5	24.5	3.8	76.9
Domicile: UK	3.5	15.4	4.0	78.7
Domicile: Other EU	5.0	13.3	8.8	67.2
Domicile: Non-EU	10.6	61.6	9.7	61.6
Ethnicity: Asian	5.2	25.5	5.9	68.9
Ethnicity: BME	6.1	47.9	6.8	68.8
Ethnicity: Black	6.4	32.4	6.7	69.7
Ethnicity: White	2.7	34.7	3.0	80.1
Ethnicity: Other	5.8	29.4	6.2	68.6
IMD: Disadvantaged	6.7	54.6	5.8	73.6
IMD: Not disadvantaged	5.7	42.1	4.9	80.6
POLAR: Disadvantaged	4.2	18.2	3.7	82.5
POLAR: Not disadvantaged	3.5	28.9	3.5	76.4
Level of study: First degree	1.9	16.6	2.6	83.4
Level of study: Other undergraduate	5.7	59.5	6.1	77.6
Level of study: PG-UG border	8.6	34.5	11.8	99.5
Welsh medium: 0-5	25.2	32.0	25.2	32.3
Welsh medium: 5-40	72.5	83.8	77.4	80.2
Welsh medium: Over 40	49.2	74.2	66.0	71.6

Notes:

- *Each observational unit is a provider-metric pair*
- *BME: black and minority ethnic; EU: European Union; IMD: Index of Multiple Deprivation; PG-UG: postgraduate-undergraduate; POLAR: Participation of Local Areas*
- *Descriptions of the split categories can be found in the Teaching Excellence and Student Outcomes Framework Specification, October 2017 (DfE, 2017a)*

5.3 Data analysis of pilot subject-level metrics

Whilst the majority of the empirical analysis in this evaluation has focussed on the TEF Year Three and Year Four provider-level data, this section assesses the robustness of the inferences (z-scores, flags, and Step 1a starting points) obtained from the 2018/19 pilot subject-level data. These data were supplied to the reviewers in an anonymised format by OfS, and conform to the 2018 method described in Section 2 of this report (including the use of nine rather than six metrics).

This section presents simple descriptive analyses of sample sizes and the prevalence of non-reportable metrics, pertaining to both core and split metrics for the pilot subject-level data. Only those 34 subjects that were assessed in the 2018/19 pilot, as described in the published TEF documentation (OfS, 2018.44), are included in the analysis.

In terms of the core metrics at provider-level, 5.8% of provider-metrics have a sample size of fewer than 30 students (the 'textbook' minimum sample size required for the Central Limit Theorem to apply), and 34.2% of providers reported fewer than the full set of nine metrics. However, there is clear heterogeneity in these results when the data are disaggregated by subject (see Table 17), with some subjects exhibiting a greater tendency towards smaller sample sizes and fewer reportable metrics than others. Over 20% of samples consisted of fewer than 30 students for 'combined and general studies' (24.5%), 'sport and exercise studies' (21.0%) and 'general, applied and forensic sciences' (20.7%), with a further 16 subjects having fewer than 30 students in at least 10% of their samples. 'Combined and general studies' also exhibited the greatest percentage of providers that were not able to provide all nine metrics (70.0%), with this being true for at least half of providers for a further seven subjects.

Recommendation 24: TEF users should be advised of potential small-sample-size issues when working with the subject-level data; consider making explicit reference to subjects where this is likely to be of particular concern.

As one might expect, when the data are further disaggregated into subject-level split metrics (also in Table 17), the prevalence of small sample sizes and non-reportable metrics becomes even more pronounced. At least 20% of samples consist of fewer than 30 students for all but 8 of the 34 analysed subjects, while the subject with the greatest prevalence of small samples, 'general, applied and forensic sciences', now consists of fewer than 30 students for over one-third (36.9%) of its samples. For all subjects, the vast majority of providers failed to provide all nine metrics for their split data. In the best case, 27.9% of provider-splits had all nine metrics reportable for 'nursing and midwifery'; conversely, just 4.1% of provider-splits had all nine metrics reportable for 'veterinary sciences'.

Non-reportable metrics for the pilot subject-level TEF are explored further in Annex C (Tables 22a and 22b), where the distribution of providers reporting 0, 1, 2, …, 9 metrics are reported in full. Among providers not reporting all nine core metrics, the majority reported at least seven metrics for 23 out of 34 subjects, rising to 30 out of 34 subjects for at least six metrics. However, among provider-split combinations not reporting all nine split metrics, the majority reported fewer than four metrics for 31 out of 34 subjects, while the percentage reporting no split metrics at all ranged from 17.4% ('medicine and dentistry') to 41.0% ('materials and technology').

The results reported above in relation to Table 17 and Annex C call into question the robustness of the inferences (z-scores, flags, Step 1a starting points) obtained from the subject-level split metrics, and reinforce our earlier recommendation from Section 5.2 with respect to split metrics: that consideration should be given to removing splits with a high prevalence of small sample sizes, or at least collapsing their categories. It should be noted that our simple analysis of the pilot subject-level data is purely descriptive in nature, and is limited to the quantification of the extent to which small samples and non-reportable metrics are present in the data, rather than assessing their impact. We have not, for example, considered the potential impact on the Step 1a outcomes, nor have we attempted to repeat our analysis of multiple comparisons from Section 3.4.6 (though it obviously follows that disaggregation of the data from provider-level to subject-level, possibly with splits on top of this, means that an even greater number of tests will need to be conducted and subsequently the potential for spurious flagging will further increase).

Table 17. Descriptive statistics on sample sizes and number of reportable metrics, 2018/19 subject-level TEF pilot

Subject	Core metrics		Split metrics	
	% samples with <30 students	% units* with <9 reportable metrics	% samples with <30 students	% units* with <9 reportable metrics
All subjects (provider-level)	**5.8**	**34.2**	**13.1**	**64.9**
Agriculture, food and related studies	17.0	45.6	30.4	90.4
Allied health	18.8	50.0	27.5	88.1
Architecture, building and planning	14.9	57.0	23.3	89.6
Biosciences	7.6	29.4	22.6	80.6
Business and management	11.2	43.1	17.3	80.4
Chemistry	5.7	24.0	26.0	89.7
Combined and general studies	24.5	70.0	29.2	92.4
Computing	19.7	35.9	28.5	84.0
Creative arts and design	11.5	32.0	24.3	81.2
Economics	7.8	29.0	24.0	89.8
Education and teaching	16.6	53.2	23.6	88.4
Engineering	10.8	55.4	18.4	87.6
English studies	4.5	12.8	20.3	78.2
General, applied and forensic sciences	20.7	44.7	36.9	92.1
Geography, earth and environmental studies	4.5	13.4	19.6	83.6
Health and social care	16.2	49.2	30.2	89.2
History and archaeology	3.4	12.0	19.6	80.3
Languages and area studies	8.8	31.1	26.3	88.1
Law	6.7	18.2	17.2	74.6
Materials and technology	14.0	65.2	35.3	93.9
Mathematical sciences	6.1	22.7	21.2	86.1
Media, journalism and communications	9.1	29.0	25.2	81.5
Medical sciences	4.9	31.4	21.4	87.3
Medicine and dentistry	2.4	29.0	10.9	89.3
Nursing and midwifery	4.2	25.4	9.5	72.1
Performing arts	14.6	39.2	25.4	86.7
Pharmacology, toxicology and pharmacy	13.0	38.9	24.4	89.2
Philosophy and religious studies	11.5	52.7	34.4	92.5
Physics and astronomy	12.9	33.3	23.1	92.7
Politics	12.8	42.9	28.7	89.1
Psychology	6.0	18.3	16.3	75.5
Sociology, social policy and anthropology	9.5	35.9	20.3	81.2
Sport and exercise sciences	21.0	40.0	29.2	86.6
Veterinary sciences	15.8	50.0	30.3	95.9

** Each observational unit is a provider for core metrics and a provider-split combination for split metrics*

6. Statistical infrastructure

In this section, we consider two wider topics – harmonisation and classifications – that have an impact on the quality of the data collected and its use in the Teaching Excellence and Student Outcomes Framework (TEF).

6.1 Harmonisation

Harmonisation is an important means of improving comparability and coherence of statistics. It includes the development of recommended questions and definitions that can be used in data collection across government. The Government Statistical Service (GSS) has published harmonised principles.

This section of the evaluation examines the degree of comparability between the GSS harmonised principles and the definitions used by the Higher Education Statistics Agency (HESA) to collect data about students in the Student Record. The data collected in this exercise are then linked to other data (survey responses, for example) and thus can form an input into TEF.

We have considered the HESA variables for which harmonised principles have previously been developed. Of these:

- Disability, Ethnicity and Sex feature as core-metric splits or contextual data in TEF
- Gender identity does not feature in name TEF, but the HESA definition of Sex has some overlap with it
- Caring responsibilities, National identity, Religion and Sexual orientation do not currently feature in TEF, but plausibly could at some future time. As such, we include our analysis of these topics as Annex D.

6.1.1 Disability

HESA provides a lot of guidance on how the question should be answered and then asks a very simple question under the assumption that all of the guidance has been read and understood. This differs from the recommended question in the harmonised principle, which includes more detail on what is classed as a disability in the question wording.

If the HESA question is fully understood by the respondent then it could be close enough to the harmonised question; however, this includes multiple assumptions about how respondents approach the HESA question:

- that they read and understand the full guidance
- that they understand that the impairment must limit their day-to-day activities
- that they understand the time requirement for an impairment to be considered ‘long-term’.

Recommendation 25: To ensure harmonisation across government data, we recommend adoption of the GSS question on disability. If not, then ensure that respondents fully understand the guidelines when answering the existing question.

6.1.2 Ethnicity

The HESA question is broadly aligned with the GSS harmonised question.

6.1.3 Gender identity

There is not currently a harmonised principle for gender identity, although this is in development. It seems likely that the current HESA variable will broadly harmonise at a categorical level with any recommended harmonised question in future. However, we make two suggested improvements.

Suggestion 1. Regarding the current HESA specification, the name of this variable could be viewed as misleading, as gender would be 'man', 'woman', etc. The data held here, labelled 'yes'/'no' identifies transgender status. It is the right thing for HESA to be collecting in terms of monitoring equalities impact, but it would be advisable to change the variable name, to avoid confusion with actual gender variables.

Suggestion 2. This information is very sensitive, and people with gender-recognition certificates do not have to disclose this information. So, we would advise including 'prefer not to say' as a standard option, as recommended by the Equality Challenge Unit. The Census question will be voluntary, either through a 'prefer not to say' option, or via a voluntary label.

Recommendation 26: To improve comparability on gender identity, and to consider alongside the definition of sex:

- re-label the variable to 'transgender status' to avoid confusion
- ensure 'prefer not to say' is included as a standard option.

6.1.4 Sex

The GSS harmonised principle currently specifies sex as male/female, and as published in the white paper, the sex question in the 2021 UK and Wales Census will remain binary. This means that the sex variable defined in the HESA information would not harmonise with the census data, GSS social surveys and many other systems as they stand.

It is not clear how the data are treated by the Education and Skills Funding Agency (ESFA) – there is an 'other' category, but the notes seem to suggest that institutions would need to provide data to the ESFA in binary form, as there is currently no non-binary legal sex. This could result in data quality issues depending on how non-binary respondents are treated when data are required in a binary format.

Assuming students are presented with male/female/other options, the resultant data are likely to be a mixture of sex and gender – as an 'other' category has been found to conflate issues of sex and gender. This introduces difficulties in measuring males and females as defined by the Equality Act, and reduces the harmonisation of data.

The information does not include information on exactly how students should be asked, so it is not clear if the data labels are being recommended as response options. But, if this response category is retained, caution should be applied to using the term 'other' to identify the non-male/female group in the actual question as it can have 'othering' implications.

Recommendations 27: Further consideration should be given to how non-binary data on respondents' sex are treated, and the implications for data quality when binary data on sex is required for reporting purposes.

6.2 Classifications

In this section, we comment upon two classifications used in TEF: the Standard Occupational Classification (SOC), and the Common Aggregation Hierarchy (CAH) of academic subjects.

6.2.1 Standard Occupational Classification (SOC)

SOC is the GSS classification of occupations, and is revised every 10 years in line with the UK population census. The current version is SOC 2010, although the coding index that supports it is updated more frequently with addition of new jobs titles. SOC 2010 is a four-level, hierarchical classification, and is largely comparable with the International Standard Classification of Occupations (ISCO). SOC is also an input to the National Statistics Socio-Economic Classification (NS-SEC), which can provide a classification for everyone, not just those in employment.

The Destination of Leavers from Higher Education (DLHE) survey and Graduate Outcome (GO) survey both code the occupation of graduates in employment to SOC, which is re-assuring as it is the harmonised standard, and thus allows greater comparability with other occupation-based statistics.

In this evaluation we have not examined the coding process itself, an operation that requires sufficient information to be provided by the respondent to allow for a detailed (4-digit) SOC code to be assigned accurately.

The SOC code is also used in the definition of the 'Highly skilled employment' metric, in which 'highly-skilled' is defined as having a SOC code in major groups 1, 2 or 3; this seems a reasonable definition.

We note that a revision to the classification, SOC 2020, is due to come into use relatively soon (GSS).

Recommendation 28: Plans and preparations should be made to handle the discontinuity caused by the forthcoming transition from SOC 2010 to SOC 2020.

6.2.2. Common Aggregation Hierarchy (CAH)

Naturally, for subject-level TEF evaluations, the classification of subjects that is used is important. Three classifications of subjects are discussed by HESA:

- Joint Academic Coding System (JACS)
- Higher Education Classification of Subjects (HECoS)
- Common Aggregation Hierarchy (CAH).

We briefly examine these three classifications and the relationships between them, but note that is it CAH that is used in TEF - and should be in use across statistics in the higher-education sector to aid consistency – to define subjects (subject groups) for TEF subject-level assessments.

Academic subjects are currently classified according to JACS. For example, the subject Mathematics is coded as G100. JACS is a hierarchical classification.

HECoS is a newer development, and is being introduced in time for implementation in the 2019/20 academic year. Mathematics is coded as 100403, and the code is described in the documentation as 'a close match' to G100). HECoS is not hierarchical, rather it is a simple list of subjects, which can be expanded as necessary as providers offer new subjects for study.

CAH has been developed recently 'as a bridge between JACS and HECoS', and 'to provide common groupings applicable to both'. The intention is that CAH will continue after the introduction of HECoS to provide a hierarchical structure. CAH is a three-level hierarchical classification. CAH1 (CAH Level 1) has 23 groups, CAH2 has 35 groups, and CAH3 has 167 groups. CAH Level 2 is being used for the subject-level TEF pilot.

We have looked further into the mapping of JACS and HECoS into CAH. The mapping of HECoS into CAH2 is strictly hierarchical or many-to-one: each code in HECoS is mapped into precisely one CAH2 group, and several HECoS codes can map into the same CAH2 group. In that sense, HECoS functions as the coding index for CAH. The mapping of JACS3 into CAH2 is not strictly hierarchical; rather it is many-to-many, although there is likely to be a strong correlation, with most cases being well-defined. Nonetheless, a single JAC3 code can map into more than one CAH2 code. An example illustrating how subject codes map and split, and based on the subject of Mathematics, is given in Annex E.

Our thoughts on the CAH classification itself are described below:

Comment 1. The year the classification was agreed or defined (or some other nomenclature that denotes the classification version) is not included in the classification's name. That might be worth introducing, for example 'CAH 2018', in much the same way as SOC 2010 and SIC 2007 (Standard Industrial Classification) are used in official statistics.

Comment 2. The name itself – Common Aggregation Hierarchy – doesn't actually say what it's an aggregation of (that is, academic subjects or degrees), whereas the names of SOC and SIC include the subject of the classification ('Occupational' and 'Industrial' respectively).

Comment 3. Since the mapping of JACS to HECoS is not entirely hierarchical, the change in classifications will bring a discontinuity. We have not investigated what statistics (outside TEF) are published or used by subject, but there will be discontinuities that will need to be handled and communicated in a suitable way. Many statistical processes could be affected, as well as the outputs themselves, and for TEF the most likely affected aspects could be benchmarking (subject is used to define some benchmark groups) and the subject-level ratings.

Recommendations 29: On the Common Aggregation Hierarchy:

- consider the name of CAH, perhaps adding 'of Academic Subjects' to its title to make it more self-explanatory, and a year to denote its introduction
- plans and preparations should be made to handle the discontinuity caused by the forthcoming transition from JACS to HECoS.

Beyond the CAH classification itself, we note a number of potential issues in its use, as described below:

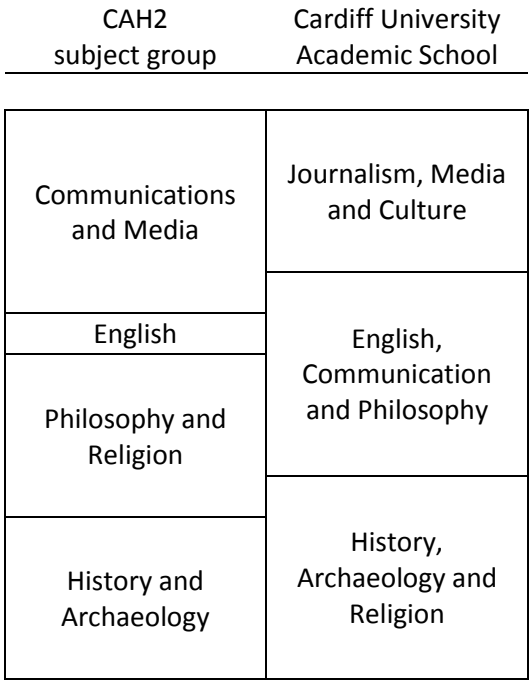
Comment 4. Our first consideration is about granularity and usefulness. Are subject-level assessments at the CAH2 level sufficiently granular to be meaningful for prospective students who would likely want to know about teaching of a specific course, rather than a broader grouping? (Of course, many taught modules would be common to a number of courses).

For example, a prospective student wanting to know about teaching excellence in relation to religious studies would look to the CAH2 group Philosophy and Religion, and would have to bear in mind that the teaching of philosophy would also be contained in the measure.

Comment 5. A related consideration is that providers' organisational structures do not always align with the structure of CAH. As an example, we consider the alignment of a selection of CAH2 subjects with the names of various Academic Schools (organisational units) found at Cardiff University (Cardiff being chosen for no particular reason). We note this is just a comparison of names - we have not investigated further the details of the courses taught by these Schools, and a full list of Cardiff University's Schools and Colleges can be found.

Figure 9 gives a schematic representation of the alignment, and is in no way 'to scale' in terms of the extent of overlap.

Figure 9. Schematic diagram of the alignment and misalignment between a small selection of CAH2 subject groups and Cardiff University Schools



That there is some misalignment between CAH2 and the various Academic Schools is obvious. If we assume that Schools have some level of autonomy and therefore that different teaching practices exist between Schools, with more similarity present for subjects taught within Schools, then the misalignment has implications for the usefulness of the TEF subject-level assessments.

For example, if the same prospective student were considering a religious studies course at Cardiff University, then the teaching would come from the School of History, Archaeology and Religion. As already noted, the relevant TEF metrics would report about CAH2 group Philosophy and Religion, but this would also contain reflections on teaching and student outcomes from the School of English, Communication and Philosophy, which may function in a different way.

Summary thoughts on the use of CAH Level 2:

We have suggested that use of the CAH2 subject-group structure may be too broad to give sufficiently useful information for prospective students. Further, it could even be misleading, as information about factors such as internal organisational structures, which may differentially affect teaching excellence and learning environments, may not be explicitly shown.

In terms of feedback for the providers, having subject-level outcomes that cannot easily be disentangled into different organisational units may not help the introduction of changes aimed at improving teaching and the student experience.

We note that National Student Survey (NSS) results are made available for subjects at the more-detailed CAH Level 3, however, which prospective students and providers alike could utilise. Obviously, at this detailed level, concerns about robustness (small sample sizes and statistical uncertainty) would be amplified in general and it may prove that very few metrics would remain reportable if TEF assessments were ever considered at this level. Nonetheless, improving access to, and information about CAH Level 3 statistics that already exist should be considered.

A further option might be the use of a 'wavy-line' type classification, whereby CAH Level 3 is used for subjects where sample sizes are sufficient for analysis, and where that is not the case, CAH Level 2 (or even Level 1) is used instead. We make no formal recommendation about this – indeed work to consider sample sizes might already have taken place in the development of CAH – and we note that there would, no doubt, be various practical implications regarding its possible use too.

7. Communication: clarity and transparency[1]

In this section we consider the ways in which the final awards and the Step 1a starting points are communicated to their audiences. We are looking for a clear description of the award given or proposed, including contextual information on benchmarking and its interpretation. We are also looking for information of sufficient quality on the methods and data used, so that users can understand the uncertainty and take that into account in their decision making.

7.1 The different audiences and their needs

In thinking about the clarity and transparency of communication on the Teaching Excellence and Student Outcomes Framework (TEF), we can think of five broad groups of users:

- The TEF assessors and Panel are seeking to make a robust judgement. They need a clear understanding of the uncertainty in the results presented to them, how to balance possibly contradictory information, and where quantitative results are more uncertain and further qualitative corroboration is needed.
- Providers will seek to understand how to get the best rating for their institution, including the presentation of the provider statement to the TEF Panel. They may be critical of their own award and decisions taken to make that award.
- The media and specialist press reporting on the higher education sector will be seeking interesting stories on exceptional or possibly controversial cases. They may be critical of the process and how it relates to controversial topics like student finance, and so may focus on issues of uncertainty or any perceived lack of transparency in decision-making.
- Students will be reflecting on their own experience of higher education, and potential students (and those advising them) will be choosing which course to apply for. Students may be critical of their own institution and seeking objective information to inform such judgements. Potential students will want to understand how different providers could meet their needs and how to make fair comparisons between providers. Both will be interested in both the quantitative information and the provider statements.
- The Universities and Colleges Admissions Service (UCAS), advising students on how to select courses, will need a clear view of how the awards are decided and what they do and do not say about providers, so they can pass this on to prospective students. In the above link, UCAS provides clear user information on interpreting TEF and are valuable partners in conveying those.

7.2 Issues in how TEF is presented

Thinking about the statistical issues presented earlier in this report, we have looked at the way in which the TEF information is provided to users. We have divided this into issues around the provision of TEF information firstly to the TEF assessors and Panel, and secondly to external users.

Recommendation 30: Consult with a broad range of users on their understanding and use of existing TEF outputs, and how they would like them communicated to be as useful as possible.

We have done this because of the TEF Panel's role in determining the final award, although there are issues common to both.

[1] This section has been contributed by the Government Statistical Service's Good Practice Team.

How the TEF information is presented to the TEF Panel

The calculations involved in assigning the Step 1a starting point (see Section 2.2) are inherently complex. For example, the use of only flagged metrics and the different impact of positive and negative metrics to the allocation of Step 1a starting point (as shown by the starting-point diagrams) may not be what people intuitively expect. This needs a clear explanation and guidance on implications for decision making. Similarly, any implications of highly correlated metrics reinforcing an outcome should be explained.

Related to the above, we noted in Section 3 that increasing the critical value used in flagging tended to push more awards to the middle (that is, from Gold or Bronze towards Silver). It follows that institutions with greater uncertainty in the metrics will similarly be more towards the centre. Knowing this, the Panel may be more inclined to put more weight on the provider statement and make a change away from the centre so, again, this needs careful explanation.

In Section 3.4.2 we discussed the statistical uncertainty and, in particular, the specification of the target population. This is crucial to understating the reliability of the results and the interaction with qualitative information in the provider statement. So, if a provider reports that this year's cohort was unusual this will, we understand, already be captured in the uncertainty in the results.

How the TEF is presented externally

We are impressed by the accessibility, clarity and transparency of the Teaching pages on the Office for Students (OfS) website. A simple web search on the term "Teaching Excellence" found this page at the second place in the listing (alongside relevant entries from the Times Higher Education Supplement and UCAS.) However, "Teaching Higher Education" and "Teaching standards higher education" did not lead to first page listing, so this suggests an external user needs some idea of the terminology to find this information easily.

The OfS Teaching landing page is clearly laid out with a prominent block linking to the 'What is the TEF' guide to TEF. This guide includes a broad explanation of what TEF is, why it is important, how to use it and how it is judged. It may be preferable to put this as the first block on the page.

The first block on the Teaching landing page is simply titled 'Teaching Excellence Awards' and leads to a page with a list of rated institutions and their awards. This may be clear to people with an understanding of the meaning of the awards, but without that background, it could easily be misinterpreted as showing only those few institutions that have been regarded as exceptional ('on the podium') for teaching. (Given the number of institutions, there is clearly more to this, but we feel there is scope for misunderstanding). We suggest that the landing page and this page both include a brief statement on interpreting the awards.

On a related point, the simple listing of providers with awards does not give the important contextual information about the benchmarking process. This could distort the decision-making process for a student. For all users, the implications of the benchmarking process for comparing (or not comparing) institutions should be made clear.

The next block, TEF data, includes a description of the data used to make assessments, and links to technical guidance for each year. The guidance includes that for providers, Panel members, and assessors, giving a strong indication of transparency about all details of the process. Similarly, the open listing of Panel members and assessors help to build trust in the system. The TEF data are provided for all providers, in a downloaded Microsoft Excel workbook showing for each provider

every metric used together with the z-score calculation, broken down by mode and year of study and a range of demographic variables.

It is impressive to see this level of openness in making available the data used in deriving the assessments. However, we could not see the provisional Step 1a starting points in the data. Whilst there is a risk that an alternative score (and more data) could confuse, transparency on these provisional awards and the implicit level of change between this and the final award would demonstrate a further commitment to transparency.

We noticed some inconsistency in the labelling of the study year. In the 'What is the TEF?' section there is a link to 'Overview of the 2018 TEF'. Elsewhere on the website, the 2018 award is referred to as 'TEF Year Four' and 'The TEF exercise 2018-19'. It would be helpful to be consistent in the terminology around the different TEF years, so users can be sure that they are looking at the latest relevant documentation and will not be confused by subtle distinctions from wording alone.

We like the training materials provided and notably the videos, catering for a wider range of learning preferences and improving accessibility. We found the videos only after looking in more detail at the supply of Year Four data. This suggests an understandable incremental build-up of these resources. It may be better to rearrange the site so that overarching materials are on, or close to, the landing page, and year-specific materials are with the associated year's data provision.

Finally, the more expert user, including those seeking to appraise or criticise the approach, should be able to find technical documents that show every step in the calculation in unambiguous detail. In principle, this should be enough to replicate the calculations with confidence.

Recommendation 31: Consider the comments given, together with user feedback, to improve the content and layout of the TEF webpages.

8. Updates from the ONS 2016 review

This section provides an update on the progress against recommendations from the 2016 Office for National Statistics (ONS) review of data sources.

8.1 Scope and objectives of the ONS Review of Sources in 2016

ONS was approached by the Department of Business, Innovation and Skills (BIS) in November 2015 with a request to carry out a review of the data sources underpinning the metrics to be used in the Teaching Excellence Framework, now known as the Teaching Excellence and Student Outcomes Framework (TEF). The three data sources examined were:

- The National Student Survey (NSS)
- The Destination of Leavers from Higher Education (DLHE) survey
- The degree of continuation of students from years 1 to 2, taken from the student record collected by the Higher Education Statistics Agency (HESA)

The objectives of the review included:

- an assessment of the quality and robustness of the sources of information
- the implications for making clear and robust determination of an institution's or course's performance against the purpose for which the metric has been chosen
- areas of improvement and the extent to which they can be addressed within the existing sources or alternative means.

An interim report was published in May 2016 (BIS, 2016a) and a final report in September 2016 (DfE, 2016a).

In the period between the review final report being published in September 2016 and the current evaluation, there have been changes which have made some of the recommendations redundant; in particular, the DLHE has been replaced by the Graduate Outcomes (GO) survey. We have not examined this new survey in any detail.

8.2 Findings of the ONS 2016 review

A small team of methodologists with a range of experience was assembled to carry out the 2016 review; they included questionnaire-design and statistical-processing experts. The data sources were assessed against well-established, internationally recognised quality standards for official statistics, including the dimensions of quality from the European Statistical System (Eurostat, undated) and the generic statistical business process model (European Commission, undated).

The NSS and the DLHE questionnaires were assessed against data collection methodology standards for questionnaire design. A few changes were recommended for the NSS and a larger number for the DLHE.

While the target population for each of the three sources was found to be clearly defined and documented, the same was not the case for TEF. A suggestion was made to consider using both a conceptual target population which specifies the ideal target population and a practical target population, which acknowledges that there are subsets of the conceptual target than cannot currently be reached. Following on from this, without a target population defined for TEF, it wasn't possible to assess the degree of under- and over-coverage from the data sources.

Although the response rates to the surveys were considered high by modern, voluntary social-statistics standards (the review reported 71% for the NSS and 79% for the DLHE), there was still scope for non-response bias. Neither survey had non-response weighting or imputation applied. Analysis carried out in the review noted a small but potentially significant level of differential response across a number of population characteristics. The review recommended further analysis of non-responders to see the degree to which they differed in characteristics from those responding.

The methodology behind benchmarking is an important component of the framework and the report recommended a review. Finally, the ONS 2016 review recommended a greater degree of user consultation going forward.

In summary, the ONS 2016 review made six recommendations:

- improve both the NSS and DLHE paper questionnaires and the on-line DLHE questionnaire to bring them up to modern questionnaire design standards
- define the target population for TEF
- determine the extent of under- and over-coverage from the data sources; modify the data sources if possible and determine weightings to account for the remaining differences
- further analysis of the characteristics of responders and non-responders should be carried out; if differences are found, weights to adjust for the differences should be applied
- carry out a methodological review of the creation and use of benchmarks
- continue to engage with data providers and users to ensure their views and concerns are captured and addressed

8.3 The approach to revisiting the 2016 recommendations

The current ONS evaluation of the statistical elements of TEF has prioritised two particular topics – the statistical methods and processes behind the metrics and benchmarking, and transparency and communication. While it is important to revisit the recommendations from the ONS 2016 review, a relatively 'light touch' has been taken. In a similar way, the methodology of the new GO survey, which is replacing the DLHE survey, hasn't been examined in any detail.

8.4 The Graduate Outcomes Survey

The DLHE survey was managed by the Higher Education Statistics Agency (HESA), which has delivered a survey of graduates since 1994/5. DLHE captured the destinations of graduates; that is, identifying what they did after leaving a provider of higher education; it was completed by students taken six months after graduation. The answers to questions on the survey contributed to two of the 2017 metrics.

A number of drivers led HESA to consider a review of the DLHE, including a need to future-proof the collection of data for an evolving labour market where workers are adopting different types of employment, the increasing use of student outcomes data and opportunities arising from the ability to link data sources (HESA, 2017). Between June 2015 and June 2017, HESA carried out a review of destination and outcomes data, called the NewDLHE Review (HESA, no date). The review was overseen by a strategic group and a working group; both groups comprised a wide range of stakeholders, including academic educations experts, careers specialists and the National Union of Students.

Two consultations were run to gather input from stakeholders. The first took place in the summer of 2016 and invited comment on the principles behind collecting outcomes data. The second invited comment on the detailed model for the collection of graduate outcomes data and ran between March and April 2017. The feedback from both reviews contributed to the development of the new survey. The GO survey contains new questions designed to capture better the wide range of student routes forward and includes a reflection on how education has affected the graduate's current situation. Optional question banks beyond the core questions allow for further information to be captured.

For the implementation phase of the GO survey, a steering group was formed to advise HESA. It comprises members from higher education providers of different sizes and types of provision across the UK, the National Union of Students, the Office for Students (OfS) and HESA and includes a senior methodologist from ONS.

To accommodate different graduation dates through the year, four cohorts are specified each with a specified 'census week'; the fourth collection period captures the majority of graduates each year. The arrangements for mode of collection have changed, from a mix of paper and electronic with telephone follow-up to electronic as the main mode with telephone follow-up. For students graduating from the 2017/18 academic year, the first census period was held in December 2018 and the last will be in September 2019; the first publication of data is expected in January 2020.

Though a detailed examination of the methodology and the development process hasn't taken place, the general approach is judged as being in line with good practice.

8.5 Department for Education response to the ONS recommendations

A formal reply to the ONS 2016 review recommendations was included in the government response to the technical consultation published in September 2016 as Annex C (DfE, 2016b). Further information was provided by a summary briefing note put together by DfE (with support from OfS) for the TEF Statistics Steering Committee in 2019 (unpublished). This evaluation received additional information from OfS and HESA.

Recommendation 1 of 2016 review: Improvement to the NSS and DLHE questionnaires

The response document states that this recommendation would be addressed as part of the normal review cycle of the surveys and the briefing note says that the NSS interface was updated by the survey contractor for the NSS in 2017.

The ONS 2016 review noted that the questionnaire was simple and clear and made only a few specific observations – these included the need for a promise of confidentiality and an estimate of the time of completion. Looking at the 2019 NSS landing page, there is a prominent confidentiality statement and an estimate of 10 minutes to complete the core questions is given in the Q&A link.

A more extensive set of observations was made for the DLHE questionnaire; however, with the replacement of the DLHE by the Graduate Outcome Survey, these observations are no longer relevant. As part of the development and testing of the GO survey, the questions and routing were examined in a cognitive testing programme run by a specialist company, IFF Research (HESA, 2018). As a result, the questions haven't been examined in this current review. This recommendation is judged to be **complete**.

Recommendation 2 of 2016 review: Definition of target population

The formal response provides links to two documents which define the target population. BIS (2016b) provides a brief description with a fuller account found in DfE (2017a). There is a necessary degree of complexity reflecting the broad range of subjects, providers and qualifications.

The target specification is given as students within the following:

Undergraduate provision leading to qualifications at levels 4, 5 and 6 for provision in England, Wales and Northern Ireland and at levels 7, 8, 9 and 10 in Scotland. For clarity, the following are in scope:

- higher and degree apprenticeships, if they include a qualification within the UK Framework for Higher Education (HE) Qualifications
- primary qualifications (or first degrees) in medicine, dentistry and veterinary science
- integrated masters degrees
- Higher National Certificates and Higher National Diplomas at levels 4 and 5.

All modes of delivery, including full- and part-time and distance, work-based and blended learning are in scope of TEF. The delivery of UK awards by overseas HE providers, or by overseas campuses of UK providers are outside the scope of TEF.

The statement of scope hasn't been explored in detail; the fact that it has been documented is taken as indicating that this recommendation is **complete**.

As a further note, previous sections of this report have identified the target population as being a super-population of students; for example, this is used to estimate NSS confidence intervals.

Recommendation 3 of 2016 review: Determine the extent of under and over-coverage; modify the data sources where possible and determine weightings to account for remaining differences

There are students in the TEF target population who are not included in the data sources and students in the data sources but not in the metrics. For example, students on short courses (1 full-time equivalence or lower (OfS, 2018.44a, paragraph 8)) are included in TEF but are not in the eligibility criteria for the NSS; non-UK students are excluded from the 'Continuation' metric because of issues with prior attainment data. For employment metrics, non-UK students are also excluded from the 'Sustained employment or further study' / 'Above median earnings threshold or in higher study' metrics, as the LEO dataset covers UK tax payers only.

It is a complex picture. The guiding principle is that the metrics aim to include as much of the TEF target population as the data sources allow. The formal response document notes that an OfS development programme starting in 2019 will consider students in the TEF target population not included in the data sources and metric definitions.

Without detailed study, it is not possible to understand the effects of the limitations of the metrics relative to the TEF target population. However, the calculation of metrics represents the first step in the assessment process and assessors are aware of their limitations.

The formal response notes that over-coverage in the datasets is accounted for by excluding any students and providers who are out of scope of TEF. For under-coverage, there are two possibilities:

- a provider in scope may not participate
- a provider and/or students in scope may not be included because of poor data quality.

Providers who do not participate, or whose submission is of sufficiently poor quality to be excluded, do not receive a TEF rating. This affects the provider concerned, but there isn't an immediate effect on other providers at provider-level, but may do at subject-level should the main provider of a subject not participate. At subject-level, there would also be a potential impact for students where only some providers have metric data. Non-response is addressed in the next recommendation.

Work on this recommendation is **expected to continue.**

Recommendation 4 of 2016 review: Investigation of the possible effects of non-response bias

The response rates for the NSS and DLHE could be considered 'high' for a social survey; the overall response rate for the 2017 NSS was 68.2% (this was despite a boycott of the survey championed by the National Union of Students who objected to the use of the data in the process to decide whether fees could be raised). There is, as expected, variation in the response rate across providers and differences when the student population is dis-aggregated into categories. Should the characteristics of non-responders differ from those responding, then adjustments can be made to account for this; for example, through weighting. The 2016 ONS report suggested using the student record as a source of characteristics data to investigate any non-response bias.

OfS has carried out an analysis of the characteristics of responders and non-responders and calculated weights. When these were used to adjust the NSS results they found only small changes. It was concluded that the results of the NSS are not suffering from bias from non-response (OfS, 2018). It would be prudent to check for non-response basis on a regular basis.

Other study looked at an imputation approach, where the characteristics of non-responders were matched with comparable responders and the answers to questions imputed from their answers (HEFCE, 2016). Comparable responders provide a range of answers, so the imputed values were sampled from a distribution of answers and the effects on the survey results calculated; further, the consequential effects on the TEF outcomes at Step 1a were also calculated. This analysis found that there were changes in the TEF awards at Step 1a for 7 providers out of 476. The paper recommended further work before any firm decisions are taken on the use of imputation. The OfS analysis of non-response looked at the NSS; they are also planning to look at adjusting for non-response in the GO survey. Work against this recommendation is **expected to continue.**

Recommendation 5 of 2016 review: Carry out a methodological review of the creation and use of benchmarks

Following the review of data sources in 2016, ONS was asked to carry out a review of benchmarking. At the time, methodological resource at ONS was not available, so this work did not go ahead. The OfS briefing note states they are carrying out a review of benchmarking methodology with reporting expected to be in 2019. OfS also commissioned an external review of alternative approaches by Alma Economics, the report of which has since been published (OfS, 2019a). This work is **expected to continue.**

Recommendation 6 of 2016 review: Continue to engage with data providers to ensure their views are captured and addressed

Developments for TEF have been subject to consultations. The GO survey has been through two consultations with responses feeding into the development which has been overseen by a strategic group and a stakeholder group.

The main TEF development has also been subject to consultation exercises; most recently regarding subject-level TEF assessments. The on-going approach to user consultation is being assessed by the Government Statistical Service Good Practice Team and is reported in Section 7 of this report.

Clearly, user engagement is ongoing; however, the original recommendation is considered to be **complete**.

9. Comment on the appropriateness of TEF metrics

In this section we consider the Teaching Excellence and Student Outcomes Framework (TEF) metrics in the context of the 'Choosing the Right FABRIC – A Framework for Performance Information' document, produced by the National Audit Office (NAO), with the acronym FABRIC representing the properties of a good system of performance information: **F**ocused, **A**ppropriate, **B**alanced, **R**obust, **I**ntegrated, **C**ost effective. The section of our evaluation is based on findings reported in an interval evaluation: ONS (2019b).

Each of the nine 2018-method TEF core metrics was considered against the eight criteria found in the FABRIC (NAO), namely:

1. Relevant to what the organisation is aiming to achieve
2. Able to avoid perverse incentives - not encourage unwanted or wasteful behavior
3. Attributable: the activity measured must be capable of being influenced by actions which can be attributed to the organisation, and it should be clear where accountability lies
4. Well-defined: with a clear, unambiguous definition so that data will be collected consistently, and the measure is easy to understand and use
5. Timely, producing data frequently enough to track progress, and quickly enough for the data to still be useful
6. Reliable: accurate enough for its intended use, and responsive to change
7. Comparable: with either past periods or similar programmes elsewhere
8. Verifiable: with clear documentation behind it, so that the processes which produce the measure can be validated.

We consider some hypothetical situations that could occur as a result.

Our main concern on the appropriateness of the TEF core metrics lies around avoiding perverse incentives. In our opinion, it is likely that providers will devise strategies that treat the metrics like targets, which could lead to behaviours that are not in the best interest of students.

A likely target could be the 'Continuation' metric as it has the highest weighting. Providers may encourage students to finish courses, change courses or go to another provider even if it is not in the interest of the student financially or personally.

Providers could take the view that students from higher socio-economic backgrounds are likely to have better labour market outcomes (Longitudinal Education Outcomes (LEO) metrics) controlling for educational attainment, and could bias their admission processes against students of lower socio-economic background. This would be in direct contradiction of the government aims of social mobility through the higher education (HE) sector. Three core metrics could be gamed in this way.

Providers may also devise strategies to boost their scores on the National Student Survey (NSS) which are not attributable to what they were set to measure, namely teaching and support provided. Five core metrics could potentially be gamed this way.

An overview of our assessment (green, amber or red) of each core metric compared against each FABRIC criterion is given in the Table 18. Note that this assessment is not an exact science, but based on professional judgement.

Table 18. Assessment of TEF core metrics against 'Choosing the Right FABRIC' criteria

'FABRIC' criteria → / TEF metric ↓	1. Relevance to organisation	2. Avoid perverse incentives	3. Attributable	4. Well-defined	5. Timely	6. Reliable	7. Comparable	8. Verifiable
The teaching on my course (NSS)	G	A	G	G	G	G	G	G
Assessment and feedback (NSS)	G	A	G	G	G	G	G	G
Student voice (NSS)	G	A	G	G	G	G	G	G
Academic support (NSS)	G	A	A	G	G	G	G	G
Learning and resources (NSS)	G	A	G	G	G	G	G	G
Continuation (HESA)	G	A	A	G	G	G	G	G
Highly skilled emp. or higher study (DLHE)	G	A	A	A	G	A	A	G
Sustained emp. or further study (LEO)	G	A	A	G	G	G	G	G
Above med. earn. thresh. or higher study (LEO)	G	A	A	G	G	G	G	G

G: Green; A: Amber; R: Red

We provide further comments on the TEF core metrics below, to help explain our assessments in Table 18. We don't comment on all criteria for all metrics, but mainly where there are concerns or where we think a 'Green' assessment might be questioned.

List of strengths and further considerations when comparing TEF metrics with the criteria of the 'Choosing the Right FABRIC' document

Core metric	Strengths	For further consideration
NSS metrics (grouped)	• NSS is easy-to-use and understand and provides data on student perceptions of topics related to teaching • Annual frequency is enough to track progress and not be subject to within-year variations • NSS has been responsive to changes, adding and changing questions • Provides a recent view and (mainly) comparable	• Could encourage the provider to lower grade boundaries to encourage students to give more positive responses on the survey • There could exist an incentive to provide less critical feedback to students • The 'Student voice' metric is more for the provider than for the benefit of the current cohort of students • Students may compare support they received with that received by peers, rather than on a what's-needed basis • Could encourage providers to supply resources that may not be required and could

	estimates over time, except for small changes to questions	be unused by students; this would not be cost effective • We couldn't find explanations for all changes to NSS questions • We note that these metrics, as measures of student perceptions, are not direct measures of teaching quality
Continuation (HESA)	• Clear and unchanged definition: comparable over time	• Might providers supply financial incentives to encourage continuation • Students moving to different providers seems like 'gaming': it might be hiding some issue at the initial provider • Not entirely in provider's control to change this, but no doubt can influence
Highly skilled employment or higher study (DLHE)	• There is greater clarity in the new definition, including that for 'higher study'	• Omission of particular career paths, in which the necessary next qualification is not regarded as 'higher' than the previous one • Metric is different from the 2017-method, similarly named metric and may not be comparable • Employment outcomes not wholly attributable to provider, much will depend on the individual student; likewise for other 'employment' metrics
Sustained employment or further study (LEO)	• Three-year time lag between graduation and data: while this may seem untimely, there is a need to wait to determine the effect of study on employment • Data still provide a useful measure, assuming the provider changes only slowly • Comparable across time	• Uses a very broad definition of employment (one day of employment per month for five out of six months) • No incentive for providers to encourage good-quality employment • A graduate's employment cannot be attributed alone to the provider
Above median earnings threshold or in higher study (LEO)	• Similar comments on timeliness to 'Sustained employment …' • Clear definition and consistency over time	• Providers could supply financial incentives for students to enrol in higher study, even if this is not best for the students • Difficult to determine how much the provider has contributed to the above-median earning aspect, and how much is attributable to the student

Some of the concerns we have identified could be quite difficult for TEF to address, for example that the metrics become targets and hence increase the risk of gaming. However, there are other areas that could be improved more easily, for example in the definitions of higher study or sustained employment.

For clarity, we note that our evaluation in this section has been confined only to the nine metrics existing in the 2018 method, and we have considered each in its own right. We have made no assessment of how well each metric acts as a proxy for what TEF is seeking to measure, and we have also not considered any other possible sources of data: those aspects lie outside the scope of this evaluation.

Recommendation 32: Consider the comments made on the appropriateness of the core metrics, and whether any improvements could be made.

10. Use of statistical information by the TEF assessors and Panel

In this section we consider what use is made of the Step 1a starting point information, and the other, contextual data made available to the Teaching Excellence and Student Outcomes Framework (TEF) assessors and Panel.

We first consider how the assessors use the statistical information, and then present an analysis of the differences between the formulaic Step 1a starting-point outcomes, and the final outcomes awarded after the assessors' and panel's deliberations.

10.1 Assessor interviews

Nine TEF assessors or Panel members were approached to participate in this part of our evaluation; six agreed, but only four could accommodate the tight timescales for the work. Those interviewed included academics and students, and had a range of experiences covering both the provider- and subject-level assessments. The four interviews were carried out by telephone. Each lasted about an hour and were recorded, with participants' consent, for later analysis (ONS, 2019c). The interviewing and analysis was carried out by an experienced, qualified professional from the Office for National Statistics (ONS) and followed standard guidelines to provide a flexible and robust approach for dealing with qualitative data. This section of our report presents an overview of the findings from those interviews; naturally the findings cannot necessarily be generalised, but nonetheless provide some additional insights into the use of statistical information in TEF assessments.

The assessors described the overall process at provider level. It starts with a review of the automated Step 1a starting-point outcome, looking into the detail of the z-scores and the flags. The contextual data, split metrics and the provider submission are then examined to gain a wider picture of the provider. A judgement is then made as to whether the automated Step 1a starting point should be changed. The process is carried out independently for each provider by three assessors who then meet to discuss their results and try to agree. Whether or not agreement is reached, the results are passed to a wider Panel for review and for a final decision to be made.

The assessors were clear that Step 1a is a starting point only and all stages are equally important – the assessment is a mix of metrics and judgement, and should be thought of as one holistic process. Those assessors without a statistical background were given training to be able to understand the meaning of the statistical outputs such as the z-scores. One assessor wondered whether assessors with different levels of statistical expertise might weight the Step 1a element differently; however, there is a moderation process using peer review to accommodate any individual differences. In the interviews, all four assessors felt confident in interpreting the statistics; they didn't feel the need for any further statistical information in Step 1a nor more generally in the making their assessments, though did note that provider type was an important factor in their assessments.

Some assessors felt the Step 1a process is quite arbitrary, and therefore that the following, holistic assessment is crucial. The benchmarks are sometimes regarded as unrealistic for all providers to achieve, and the borderline-rating categories are not considered especially useful.

When moving on from the Step 1a starting point, the assessors examine the contextual data and the provider submission. For providers with full information and high initial scores, the assessor will look for supporting evidence. If the Step 1a starting point, contextual data and submission present a consistent picture then the initial hypothesis is unlikely to change. The examination of the contextual data can help to understand factors that might have led to a metric being flagged at Step 1a. For example, a negative flag might be influenced by a high number of 'widening participation' students. Following a consideration of all the information, the assessor makes a judgement on the award. An

interesting observation was that the gap between Bronze and Silver is regarded by some as being larger than that between Silver and Gold. In cases of missing or non-reportable metrics (often the case for smaller providers), the qualitative information is used to provide evidence for the award.

For the provider-level assessment, the four assessors were satisfied that the overall process was fair and robust. If there were individual differences in interpretation of evidence, it was felt that that the peer-review process would account for this. The assessors were happy with the mix of quantitative and qualitative information and how it is assessed; they felt there was no need for any further automation of the process. It was noted that the overall process had taken on board feedback from assessors each year and had improved as a result.

For subject-level assessment (at pilot stage), there were concerns raised. It was felt that there was a lower level of scrutiny than at provider-level, that the classification of subjects could be problematic, that greater evidence would be needed from providers who offered many combined degrees and whether sufficient data were available.

10.2 Empirical analysis of differences between the Step 1a starting point and the final TEF awards

In this section we summarise the observed differences between the Step 1a starting point and the final TEF awards based on TEF Year Three (the most recent year for which the final awards have been published), and explore the characteristics of the providers that had their award changed. This analysis is intended to be purely descriptive, and supplements the findings of the qualitative research outlined above.

It should be noted that our derived TEF Year Three starting points do not perfectly match the confirmed starting points supplied to us by the Office for Students (OfS); we derived 19 Bronze, 62 Silver and 5 Gold starting points, compared with confirmed starting points of 22 Bronze, 58 Silver and 6 Gold. This is likely caused by differences between the published methodology and the way in which the process is implemented in practice (for example, imputation of flags for non-reportable core metrics using individual component years, as discussed in Section 3.8).

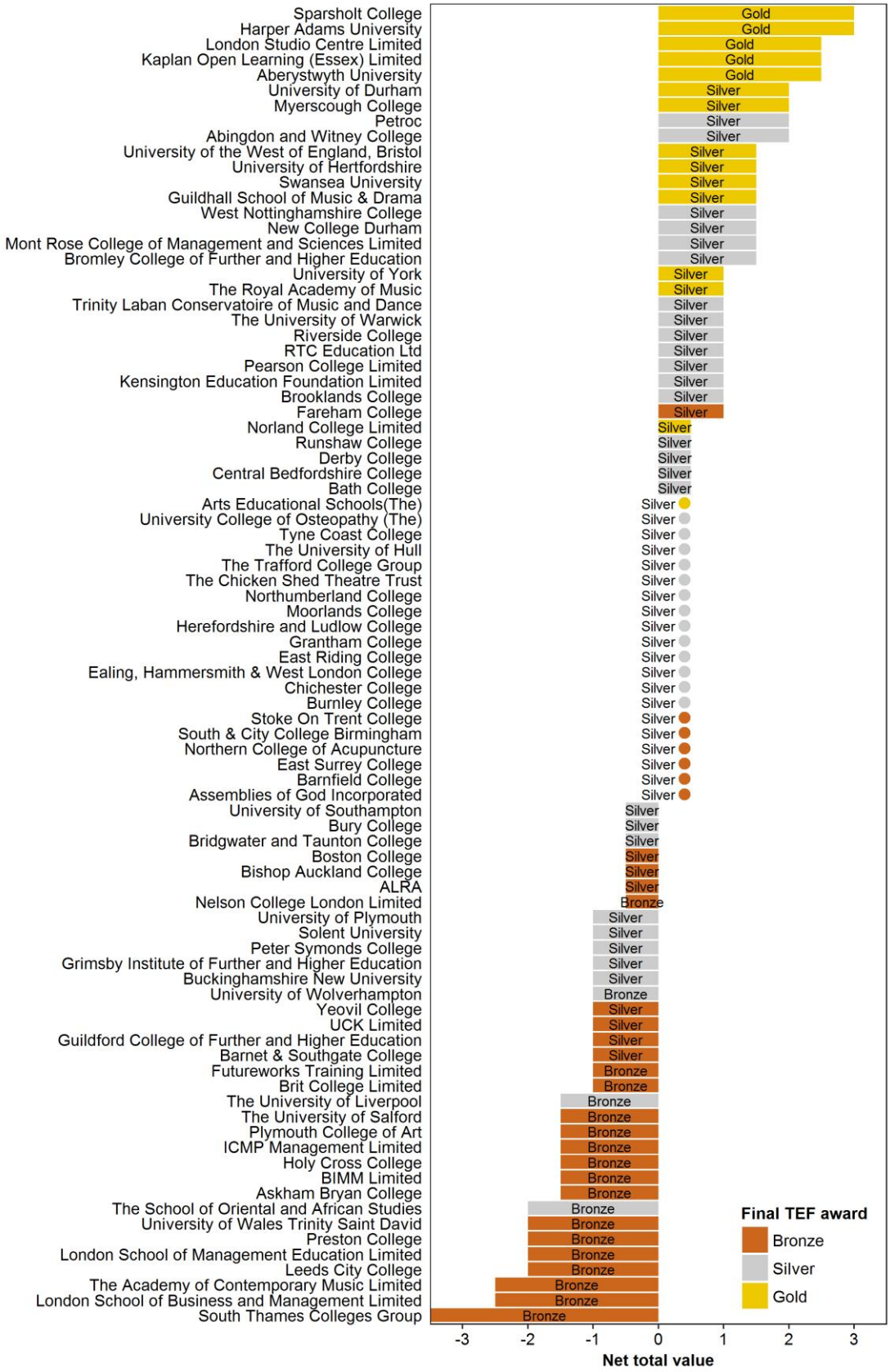
Of the 86 providers assessed, 19 achieved a Bronze at the Step 1a starting point, 62 achieved a silver, and 5 achieved a Gold; at the final award stage, these frequencies had been modified to 30, 41 and 15, respectively (see Table 19). When considering the direction of change, 13 providers' awards were upgraded (10 of which were Silver to Gold), 14 providers' awards were downgraded (all Silver to Bronze), and 59 providers' awards were unchanged. There were no instances of a provider having their award changed by more than one category (either Bronze to Gold or vice-versa).

Table 19. Step 1a starting points and the final TEF awards, TEF Year Three

	Final TEF award		
Step 1a starting point	**Gold**	**Silver**	**Bronze**
Gold	5	0	0
Silver	10	38	14
Bronze	0	3	16

The results above are visualised in Figure 10. The colours of the bars represent the providers' final awards, while the labels within the bars represent the providers' Step 1a starting point. The providers are ordered according to their net total value (the difference between the sum of their weighted positive flags and the sum of their weighted negative flags).

Figure 10. Step 1a starting points and final TEF awards, TEF Year Three, ordered by net total value



Note: figure inspired by Wonkhe (2018)

Providers that had their award downgraded between the Step 1a starting point and the end of the TEF process tended to have a low net total value at Step 1a in absolute terms: all 14 providers had a score between –1.0 and +1.0, and nine of these had a score between –0.5 and +0.5. However, this association between change in award and low absolute net total value was less apparent for providers that had their award upgraded: these 13 providers had Step 1a net total values ranging from –2.0 to +2.0, and with eight of these providers having a net total value between +1.0 and +2.0.

We note that all 14 providers that had their awards downgraded between Step 1a and the end of the TEF process were alternative providers (APs) or further education colleges (FECs) based on the Higher Education Statistics Agency (HESA) classification of higher education providers (HESA, undated); none of the 22 higher education institutions (HEIs) participating in TEF Year Three had its award downgraded (see Table 20). Meanwhile, 41% of HEIs had their awards upgraded between Step 1a and the end of the process, while this was true for just 6% of the 64 non-HEIs participating in TEF Year Three.

Table 20. Changes to TEF Year Three awards from Step 1a to the end of the process, stratified by provider type

Step 1a starting point	Change to award	Higher education institutions *n (% of Step 1a category)*		Alternative providers and further education colleges *n (% of Step 1a category)*	
Gold	Upgraded ↑	not applicable	not applicable	not applicable	not applicable
	No change =	2	(100%)	3	(100%)
	Downgraded ↓	0	(0%)	0	(0%)
Silver	Upgraded ↑	6	(43%)	4	(8%)
	No change =	8	(57%)	30	(63%)
	Downgraded ↓	0	(0%)	14	(29%)
Bronze	Upgraded ↑	3	(50%)	0	(0%)
	No change =	3	(50%)	13	(100%)
	Downgraded ↓	not applicable	not applicable	not applicable	not applicable
Overall	**Upgraded ↑**	**9**	**(41%)**	**4**	**(6%)**
	No change =	**13**	**(59%)**	**46**	**(72%)**
	Downgraded ↓	**0**	**(0%)**	**14**	**(22%)**
	Total	**22**		**64**	

We used logistic regression to investigate associations between the TEF provider-level contextual characteristics and the propensity for an award to be downgraded (Table 21a) or upgraded (Table 21b) between Step 1a and the end of the TEF Year Three process. The probability of a provider being downgraded from Gold to Silver or from Silver to Bronze was significantly negatively correlated with the proportion of students at the provider that are white (p-value = 0.025) and significantly positively correlated with the proportion of students at the provider that are UK domiciled (p-value = 0.014). The probability of a provider being upgraded from Bronze to Silver or from Silver to Gold was significantly negatively correlated with the proportion of students at the provider that are male (p-value = 0.041). We remind the reader that these results are suggestive of statistical associations, and we do not imply any causal mechanism in terms of the TEF Panel's decisions.

The preceding analysis is intended to be descriptive in nature and supports our evaluation concerning changes in TEF awards; our aim is not to construct a model that can be used to predict changes in TEF awards in some broader sense. Accordingly, we restricted the range of covariates included in the models to simple summaries of the published TEF contextual data and have not conducted a full assessment of model performance. However, some headline model diagnostics are reported in the footnotes to Tables 21a and 21b. Both models are significantly different from the corresponding intercept-only models, and neither model exhibits evidence of lack-of-fit according to the Hosmer-Lemeshow test. Both models also exhibit reasonable discriminatory power, with areas under the receiver operating characteristic (ROC) curve in excess of 90%.

Table 21a. Logistic regression output: award downgraded between Step 1a and the end of the TEF Year Three process

Variable	Coefficient	Std. err.	z-score	p-value
Intercept	-29.044	15.204	-1.910	0.056
Absolute net total value	-1.228	1.080	-1.137	0.256
Headcount	-0.004	0.002	-1.671	0.095
% students with first degree	-0.026	0.027	-0.974	0.330
% students aged < 21 years	-0.076	0.069	-1.111	0.267
% students white	-0.151	0.067	-2.249	0.025
% students male	0.019	0.041	0.475	0.635
% students disabled	-0.049	0.128	-0.384	0.701
% students with high entry tariff	0.131	0.115	1.143	0.253
% students UK domiciled	0.507	0.207	2.447	0.014
% students local	-0.111	0.060	-1.850	0.064
% students in POLAR quintile 1	0.316	0.215	1.469	0.142
% students in IMD quintile 1	0.004	0.071	0.051	0.959

Notes:

- *Model estimated on 67 providers with a Gold or Silver starting point at Step 1a*
- *IMD: Index of Multiple Deprivation; POLAR: Participation of Local Areas*
- *Descriptions of contextual variables can be found in the Teaching Excellence and Student Outcomes Framework Specification, October 2017 (DfE, 2017a)*
- *Selected model diagnostics:*
  - *Likelihood ratio test vs. null model: p-value < 0.001*
  - *Hosmer-Lemeshow test: p-value = 0.746*
  - *Nagelkerke R^2: 0.622*
  - *Area under the receiver operating characteristic curve: 0.923*

Table 21b. Logistic regression output: award upgraded between Step 1a and the end of the TEF Year Three process

Variable	Coefficient	Std. err.	z-score	p-value
Intercept	-12.884	13.559	-0.950	0.342
Absolute net total value	2.286	1.443	1.585	0.113
Headcount	0.00005	0.0002	0.281	0.779
% students with first degree	0.009	0.036	0.265	0.791
% students aged < 21 years	0.253	0.143	1.773	0.076
% students white	0.141	0.090	1.564	0.118
% students male	-0.182	0.089	-2.041	0.041
% students disabled	-0.268	0.222	-1.210	0.226
% students with high entry tariff	-0.090	0.097	-0.929	0.353
% students UK domiciled	-0.063	0.093	-0.681	0.496
% students local	-0.118	0.075	-1.582	0.114
% students in POLAR quintile 1	-0.457	0.587	-0.778	0.437
% students in IMD quintile 1	0.365	0.218	1.677	0.093

Notes:

- *Model estimated on 81 providers with a Silver or Bronze starting point at Step 1a*
- *IMD: Index of Multiple Deprivation; POLAR: Participation of Local Areas*
- *Descriptions of contextual variables can be found in the Teaching Excellence and Student Outcomes Framework Specification, October 2017 (DfE, 2017a)*
- *Selected model diagnostics:*
  - *Likelihood ratio test vs. null model: p-value < 0.001*
  - *Hosmer-Lemeshow test: p-value = 0.987*
  - *Nagelkerke R^2: 0.772*
  - *Area under the receiver operating characteristic curve: 0.977*

11. Conclusions and next steps

This section aims to draw together our evaluation's findings and recommendations, and presents them according to various themes. Note that we only identify here new, formal recommendations that have not been made elsewhere in the report; these are ones that draw on evidence from more than one part of our evaluation.

Overview

This report has considered many of the statistical aspects of the Teaching Excellence and Student Outcomes Framework (TEF). It has been conducted over a short, but intense period by the Methodology Advisory Service (MAS, based at the Office for National Statistics (ONS)) of the Government Statistical Service (GSS) and contains descriptions of the TEF process, comments on its methods, analytical research into its outcomes, and suggestions for developments.

The performance of the higher education sector, with respect to its teaching and the outcomes of students is a very complex, diverse and challenging entity to measure, and there are very many options for doing so. TEF attempts to do this, its aims seem well-founded, and it is clearly now becoming established.

The TEF assessments are based on a wide variety of data-driven metrics, processes and rules, combined with other information via a subjective, but holistic assessment by trained assessors. That TEF now has a suite of measures and data, from numerous sources, associated with teaching excellence and student outcomes brought together in one framework is commendable.

At the provider-level, the Step 1b and Step 2 process – that involving the expert judgement of assessors and the TEF Panel – seems robust and takes a holistic approach; the assessors take their role very seriously, and apply professional skills and expertise to areas that statistics alone cannot capture. The subject-level assessments trialled in the Year Four pilot seem less robust, however, as evidenced by smaller sample sizes, assessors' comments, and considerations around the usefulness of the subject classification used. These issues are likely to remain, but other improvements (if made) around communication and methods more generally, may provide some mitigation.

Recommendation 33: Review the pilot run in TEF Year Four, considering the usefulness of the subject-level ratings given the methods and data that support them.

The methods in Step 1a, which are entirely formulaic and data-driven, have clearly been developed, and continue to be developed, with diligence and expertise. However, there are some aspects of the process we have identified that do not serve the assessment well, as they have the potential to produce inconsistent or undesirable results; these may be mitigated by having assessors examine the results in more detail in later steps, but there is also the potential for them to be overlooked. We make several recommendations for the methods to be developed further.

More widely, transparency and communication of the statistical aspects of TEF should also be improved. This includes clearer communication of the fact that statistical uncertainty exists in the data and outputs, and statements about the impact this has on use of the data, and better guidance on how TEF statistical outputs should (and shouldn't) be used.

Step 1a processes

Our study of the Step 1a, formulaic rules that deliver a single starting-point for each provider or subject have identified a number of undesirable features. As a result, it would be possible for similar providers to receive very different outputs in Step 1a.

The first of these features is the binary nature of flags, in which a metric receives all or nothing of the weight of the metric depending on which side of an arbitrary threshold an indicator and the z-score lie. The use of flags – if correctly adjusted to accommodate possibly many multiple tests or comparisons, which would decrease the number of flags showing – can still provide a useful, visual indication of a statistically significant and meaningful result. However, we believe their use in providing a contribution to the Step 1a starting-point calculation needs further consideration, so as to allow proportions of the metric's weight to be realised, rather than the all-or-nothing approach.

The starting-point diagrams, used to determine the starting-point category given a provider, or subject's, positive and negative total flag values is somewhat arbitrary. It can lead to almost identical providers receiving very different Step 1a outcomes, and the 2018-method borderline categories don't seem to function particularly effectively (a view also echoed by assessors we interviewed). The starting-point diagrams also penalise negative flags more severely than positive flags contribute, despite the definitions of the metric weights and flags being symmetric around differences of zero; we are not sure whether this feature was intended, but it accords with an assessor's comment that the gap between Bronze and Silver is regarded as wider than the gap between Silver and Gold.

We recommend the introduction of a net total value measure, in which the total value of negative flags is subtracted from the total value of positive flags. We see this as an intuitive and more transparent measure, but note it and the current starting-point diagrams are inconsistent. Such a measure could be adapted to accommodate a scaling of the negative values (if desired), and the presentation of the results could be changed to be on a continuous, colour-graduated scale. Again, there is more work required to research and develop the proposal.

The choice of weights for the metrics seems somewhat arbitrary, although changes to address some identified issues have been made over time. Naturally, and assuming a single, combined measure is still desired, one 'official' set of weights is required. However, it might be useful to develop a tool that allows users to input alternative weights that reflect their own views on which metrics are the most important, and to create new, Step 1a output measures on that basis. We comment briefly on other options concerning the calculation of an overall measure, including analogous, non-benchmarked versions for comparison, and the diversity of the component metrics, which may be hidden by the use of a combined measure.

Benchmarking is clearly a contentious issue, which has both supporters and opponents. Its aims are reasonable, it has clearly been well-researched, and is an arguably more transparent alternative to an equivalent model-based approach. However, its biggest limitation is that the residual – the difference between a provider's indicator and its benchmark – cannot solely be attributed to the effect we wish to measure (the value-added by that provider in terms of teaching excellence or student outcome) as it also contains an effect from all the other factors that were not included in the benchmark definitions. Those include any innate differences between providers, and the communication of what benchmarking accounts for, and what it doesn't should be clearer.

Aspects such as student expectations and local-area variations (such as living costs) are largely outside the control of providers, but may have a bearing on observed metric values. These are not included as benchmark factors as they are often only relevant at provider-level (and not available at student level) or are otherwise very difficult to measure. Without accounting for such differences, the assessment of all providers together in TEF means that providers of different types (in terms of the sorts of factors discussed) contribute to each others' benchmarks, and make the TEF outcome more difficult to interpret as the residual is some mix of what TEF seeks to measure and what it would wish to control for.

An option to mitigate this would be to carry out TEF assessments by groups of providers, where those groups would need to be pre-defined. Establishing such groups would require further research and would be challenging. However, we consider there would be statistical benefit if this could be achieved, which would then need to be balanced against any relevant policy aims. In the later stages of the TEF assessment (Steps 1b and 2), our research has suggested that different types of provider may be treated differently anyway. For example, study of the process rules, combined with conversations we have had, suggest that smaller providers are more likely to get a Step 1a rating of Silver by default (because of missing metrics or small sample sizes), and our (limited) analysis suggests that alternative providers and further education colleges may be more likely to have their Step 1a starting point downgraded than higher education institutions. Interviews with assessors, albeit small in number, suggest that different types of providers sometimes require different treatment, for example by using the provider submissions to fill the gaps where there are metrics missing.

A number of assumptions are made in the Step 1a processes, and it would be useful to make these more explicit in the documentation. There is an assumption of normality of the differences between indicators and their benchmarks; this seems reasonable and has been researched, though it is difficult to test empirically in TEF. A wider assumption, and one that is not well documented, is the use of a super-population. There are many possible target populations, defined at different time points, that could be considered for use in TEF outputs, such as:

- the National Student Survey (NSS) targeting providers' current students
- the Destination of Leavers of Higher Education (DLHE) survey and the Graduate Outcomes (GO) survey targeting recent graduates
- the Longitudinal Education Outcomes (LEO) dataset containing data on employed graduates some time after graduation.

Those examples all target fixed and finite populations, and contain different students or graduates in the same TEF year. However, the calculation of z-scores and confidence intervals for the NSS assumes not a fixed and finite population, but rather a much bigger (infinite) population from which the responding sample is just one possible realisation. The practical implication of this, when combined with benchmarking, is that TEF considers the mix of current students at a provider to be relatively fixed (or stable over time) in terms of their characteristics (benchmarking factors), but that the actual students who happen to be present is random. This seems a reasonable approach, especially as prospective students are probably wanting to know what providers' teaching and their own post-higher education outcomes might be like, rather than specifically what has happened to another group of (former) students.

Communication: transparency and accessibility

A lot of information about TEF is published and publicly available, including guides, documentation and data. That said, as reviewers we found it somewhat difficult to be sure we'd found all the available information, and to know that what we had found was the most up-to-date and still relevant. Of course, that no doubt partly reflects the very complicated nature of TEF and its ongoing development, but we still think there is scope to consolidate all information about TEF's statistical methods and its data and outputs.

The documents describing the methods are detailed and gave almost sufficient information to allow us to recreate the Step 1a starting points. However, there are some gaps in the method documentation that should be filled, and more detail or explanation in places would be helpful. The Office for Students (OfS) has clearly been undertaking extensive analysis on TEF data, and it would be useful for that analysis to be published, even if just as working papers. Examples include analysis of year-on-year stability and a report that suggests there is currently no appreciable non-response bias in the NSS outputs, despite there being no survey weighting applied.

We think that better and clearer guidance could be provided to help users:

- interpret appropriately the differences between indicators and benchmarks
- consider the effect of statistical uncertainty in the data
- allow for multiple comparisons and the increased chance of observing 'significant' results by chance alone
- understand the assumptions that have been made in the methods and processes, such as assumptions around normality and the use of super-populations.

Other issues

We have also considered other aspects of TEF in our report. These include:

- developments in light of the ONS 2016 review, the recommendations from which are either complete or ongoing
- the appropriateness of the TEF metrics, which we consider generally to be good, although there is a risk of gaming by providers, and not all metrics are mainly attributable to the provider
- the level of harmonisation of TEF data – and more specifically the Student Record information that informs TEF splits or is provided as contextual information; this generally seems compliant with GSS Harmonised standards
- classifications: both the Standard Occupational Classification (SOC) and classification of subjects will be revised or changed soon, and it would be prudent to plan in advance for the change and possibility of discontinuities. In addition, we question the usefulness of the Common Aggregation Hierarchy 2 (CAH2) for classifying subject-level TEF assessments, noting instances where it may be too broad or cut across organisational structures.

Summary and next steps

The data and statistical methods used in TEF provide just one part of the assessment process, and it is important to bear that in mind when considering the quality and appropriateness of the methods. Our review of the methods has found a number of areas that we believe could be improved through further development or better communication.

We suggest that any changes to methods are considered in a holistic way, as one changed aspect could render others unnecessary or require further work elsewhere. On communications, there is some analysis to publish, some gaps in documentation to fill, and an exercise to consolidate what already exists.

12. Closing remarks

This brief evaluation has examined the existing statistical methods used in the Teaching Excellence and Student Outcomes Framework (TEF) and has included a small amount of data analysis and a few interviews. With only a short period for the review, we have not been able to follow through on our suggestions for alternatives to some aspects of the methodology; instead, this work is the subject of our recommendations.

As we noted in our introduction, a great deal of effort has gone into the development of TEF to this current point and we recognise the progress that has been made in what is a difficult area – measuring complex aspects of organisational performance.

Although TEF is becoming established, its methodology is still relatively new and an independent statistical review is a standard and effective way to help to improve on what has gone before. We hope that the recommendations in this evaluation will assist in forming future development plans for TEF.

References

Cardiff University. 'College Structure' (Accessed 15 March 2019)

Department for Business, Innovation and Skills (2015). 'Fulfilling our Potential: Teaching Excellence, Social Mobility and Student Choice' November 2015. (Accessed 28 June 2019)

Department for Business, Innovation and Skills (2016a). 'Teaching Excellence Framework: Interim review of Data Sources' (Accessed 5 July 2019)

Department for Business, Innovation and Skills (2016b). 'Higher Education: Success as a knowledge economy - white paper' (Accessed 5 July 2019)

Department for Business, Innovation and Skills (2016c). 'Teaching Excellence Framework: year two and beyond. Government technical consultation response' September 2016. (Accessed 28 June 2019)

Department for Education (2016a). 'Teaching Excellence Framework: Review of Data Sources' (Accessed 5 July 2019)

Department for Education (2016b). 'Teaching Excellence Framework: year two and beyond, Government technical consultation response' (Accessed 5 July 2019)

Department for Education (2017a). 'Teaching Excellence and Student Outcomes Framework Specification' October 2017. (Accessed 14 March 2019)

Department for Education (2017b). 'Teaching Excellence and Student Outcomes Framework: analysis of metric flags' (Accessed 5 July 2019)

Department for Education (2017c). 'Teaching Excellence and Student Outcomes Framework: lessons learned from Year Two' October 2017. (Accessed 28 June 2019)

Department for Education (2018). 'Teaching Excellent and Student Outcomes Framework: Subject-Level. Technical document to support the government consultation' (Accessed 3 July 2019)

Draper, D and Gittoes, M (2004). 'Statistical Analysis of Performance Indicators in UK Higher Education'. Journal of the Royal Statistical Society. Series A, Vol. 167, No. 3, pages 449-474. Wiley.

Eurostat. 'Quality Assurance Framework of the European Statistical System' (Accessed 14 March 2019)

Eurostat. 'Quality Standards for Official Statistics' (Accessed 5 July 2019)

European Commission. 'Generic Statistical Business Process Model' (Accessed 5 July 2019)

Forster, JJ (no date). 'A note on sample size effect in the TEF flagging process' School of Mathematical Sciences, University of Southampton, UK. Unpublished

Goldstein, H and Spiegelhalter, D J (1996). 'League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance' Journal of the Royal Statistical Society. Series A, Vol. 159, No. 3, pages 385-443. Wiley

Government Statistical Service. 'Harmonised principles' (Accessed 27 March 2019)

Government Statistical Service. 'Standard Occupation Classifications (SOC)' (Accessed 25 March 2019)

Higher Education Funding Council for England (2016). 'Non-response and missing data in the Teaching Excellence Framework Metrics' TEF Working Group Paper. Unpublished

Higher Education Statistics Agency. 'Benchmarks' (applicable to tables T1 to T3, T7 and E1) (Accessed 14 March 2019)

Higher Education Statistics Agency. 'Destination of Leavers from Higher Education (DLHE) survey' (Accessed 5 July 2019)

Higher Education Statistics Agency. 'Graduate Outcomes Survey questionnaire' (Accessed 25 March 2019)

Higher Education Statistics Agency. 'Guide to the UKPIs' (Accessed 15 March 2019)

Higher Education Statistics Agency. 'Higher education providers' (Accessed 15 March 2019)

Higher Education Statistics Agency. 'NewDLHE: Destination and outcomes review' (Accessed 5 July 2019)

Higher Education Statistics Agency. 'Student record 2018/19' (Accessed 15 March 2019)

Higher Education Statistics Agency. 'The Higher Education Classification of Subjects (HECoS)' (Accessed 15 March 2019)

Higher Education Statistics Agency (2017). 'NewDLHE: The Future of Student Outcomes Data' (Accessed 5 July 2019)

Higher Education Statistics Agency (2018). 'Graduate Outcomes Cognitive Testing Technical Report' (Accessed 5 July 2019)

Holm, S (1979). 'A simple sequentially rejective multiple test procedure' Scandinavian Journal of Statistics, Vol. 6, No. 2, pages 65-70. Wiley

National Audit Office (2001). 'Choosing the Right FABRIC' (Accessed 15 March 2019)

National Student Survey (2017). 'National Student Survey 2017 Core Student Questionnaire' (Accessed 14 March 2019)

Office for Students. 'Confidence intervals for the Unistats website'. Internal paper

Office for Students. 'Get the NSS data' (Accessed 14 March 2019)

Office for Students. 'National Pupil Database' (Accessed 8 May 2019)

Office for Students. No title (subject is confidence intervals for the NSS). (Accessed 14 March 2019)

Office for Students. 'Regarding confidence intervals for NSS estimates' (Accessed 19 May 2019)

Office for Students. 'Teaching pages' (Accessed 5 July 2019)

Office for Students. 'TEF Year Four – publicly available workbooks and metrics data, but methods as per TEF Year Three (2017 method)' (Accessed 17 May 2019)

Office for Students. (2018.44). 'Teaching Excellence and Student Outcomes Framework: Subject-level pilot guide', October 2018 (Accessed 14 March 2019)
[We note an updated version has been published since: 'Teaching Excellence and Student Outcomes Framework: Subject-level pilot guide' (Accessed 28 June 2019)]

Office for Students. (2018.44a). 'Teaching Excellence and Student Outcomes Framework: Guide to subject-level pilot data', October 2018 (Accessed 14 March 2019)

Office for Students (2018). 'Non-response bias in NSS' technical paper. Unpublished

Office for Students (2019a). 'Review of benchmarking methodologies, Alma Economics for OfS' (Accessed 25 June 2019)

Office for Students (2019b). 'TEF benchmarking principles' (Accessed 28 June 2019)

Office for National Statistics. 'Personal well-being' (Accessed 22 March 2019)

Office for National Statistics (2016). 'Teaching Excellence Framework: review of data sources' (Accessed 14 March 2019)

Office for National Statistics (2019a). 'Scope document for TEF statistical evaluation'. Internal paper

Office for National Statistics (2019b). 'Review of TEF in the context of the 'Choosing the right FABRIC' document'. Internal paper

Office for National Statistics (2019c). 'Teaching Excellence Framework (TEF) Review (analysis of interviews with TEF assessors)'. Internal paper

Organisation for Economic Co-operation and Development (2008). 'Handbook on constructing composite indicators: methodology and user guide' (Accessed 5 July 2019)

Royal Statistical Society (2018). 'Response to the Teaching Excellence and Student Outcomes Framework, Subject-Level Consultation' (Accessed 24 April 2019)

Royal Statistical Society (2019). 'Submission to the Independent Review of the TEF' (Accessed 15 March 2019)

Student Loans Company. 'Student loans data' (Accessed 8 May 2019)

Unistats website (Accessed 24 April 2019)

Universities and Colleges Admissions Service. 'How to select courses' (Accessed 5 July 2019)

UK Centre for Materials Education. 'The National Student Survey: An overview of UK Material Programmes 2010' (Accessed 14 March 2019)

UK Statistics Authority (2018). 'Code of Practice for Statistics: Ensuring official statistics serve the public' Edition 2.0, February 2018. (Accessed 14 March 2019)

Wonkhe (2018). 'Fun with flags – TEF3 style' (Accessed 7 May 2019)

Glossary

This a summary of commonly used terms in this report. Where possible, we have used terminology consistent with tht found in the various TEF documentation from DfE and OfS.

2017 method: the method applied to a set of six core metrics to form the Step 1a starting point for the Year Three (academic year 2017/18) and Year Four (2018/19) provider-level TEF, as described in DfE (2017a). Note that some changes were made in year Four (OfS (2018.44)), but these don't materially change the method.

2018 method: the method applied to a set of nine core metrics to form the Starting Point for the subject-level TEF pilot (in 2018/19), as described in OfS (2018.44 and 2018.44a)).

Absolute value: also called the indicator, that is without consideration of the benchmark. In mathematical terminology, 'relative value of the indicator' would be a more precise description, as providers' indicators marked as having very high (very low) absolute values are really just those that have indicator values that are relatively high (low) in comparison with other providers' values.

Aspects of quality: the three broad areas that TEF aims to assess: Teaching quality (TQ), Learning environment (LE) and Student outcomes and learning gain (SO).

Assessors: in the assessment process, TEF assessors consider the statistical and non-statistical information presented and make recommendations to the TEF Panel.

Benchmark (denoted by *E*): an expected value for a provider's indicator if it had performed as per the sector average, given its student and subject mix.

Borderline rating: the Gold/Silver and Silver/Bronze starting-point categories available under the 2018 method.

Core metric: one of the six (2017 method) or nine (2018 method) metrics whose data form inputs to the Step 1a starting-point calculations. Other metrics – measuring further aspects of providers and their students' outcomes – form contextual data. See also split metric.

Difference (denoted by *d*): defined as the Indicator minus the Benchmark for any given provider or subject within provider. For each of the current set of metrics, this is measured in percentage points.

Flag value: the value given to a metric depending on the metric's weight and whether the metric is flagged or not. Possible flag values under the 2017 and 2018 methods are, variously, +/-0.5, +/-1.0 and +/-2.0 if the metric is flagged, and 0 if not, though sometimes the it is the absolute value that is reported separately for negatively and positively flagged metrics. In Annex B we propose modifying the calculation for the flag value, but in a way that the current definition is a special case of a more general formula.

Indicator (denoted by *p*): the value of metric for a provider or subject within provider. For each of the current set of core metrics, this is measured as a percentage and of something regarded as positive or desirable (continuation rates, for example, rather than non-continuation rates).

Initial hypothesis: strictly, this is the rating of Gold, Silver, orBronze arrived at after the completion of Step 1, that is after Step 1a (providing the starting point), Step 1b and a judgemental assessment of the information provides by Steps 1a and 1b by the assessors. However, note that the term 'Step 1a initial hypothesis' is sometimes used synonymously with starting point.

Final rating: assessment of a provider (or any of its subjects) after consideration of all the evidence by the TEF Panel; possible categories are Gold, Silver and Bronze.

Flag: a symbol, available for each metric, used to denote indicator values that are significantly and materiality different from the benchmark. Flags can be positive or negative, denoting relatively better or worse observed values of the indicator, and are created based on a set of rules primarily involving the value of the Difference and the z-score of the metric. Metrics reported without a positive flag or a negative flag are denoted as unflagged, indicating – in general terms – only a small or insignificant difference between the indicator and benchmark values of that provider's metric.

Flag value: see weight.

Metric: the concept being measured, such a 'Continuation' or 'Student voice'. See also Core metric.

Panel: in the assessment process, the TEF Panel considers and debates the recommendations made by the TEF assessors, and agrees the final Gold, Silver or Bronze ratings for each provider or subject.

Provider: an establishment, such as a university of other institution, providing higher education.

Quality: (1) see also Aspects of Quality; (2) fitness-for-purpose, often characterised by the European dimensions of statistical output quality: relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, and accessibility and clarity.

Split metric: data for a metric provided in relation to one of the data splits, a demographic sub-group of the population of interest.

Starting point: a categorical rating for each provider or subject-within-provider derived solely by formulaic means and based only on the core metrics; the starting point is delivered as Step 1a in the decision-making process. Note: the starting point should not be confused with 'initial hypothesis', which emerges after a judgemental consideration of the starting point (based on core metrics), alongside information on split metrics and absolute values (Step 1b), although it is sometimes referred to as the 'Step 1a initial hypothesis'.

Step 1a (see starting point; also compare with initial hypothesis).

Weight (denoted by w): (see also flag value). A pre-determined value assigned to a flag and metric-dependent. In the 2018 method, possible flag values are 0.5, 1.0 or 2.0, depending on the metric. Positive flags carry this value or weight as a positive number, whereas negative flags may be regarded as carrying the negative of these values (-0.5, -1.0 or -2.0), though reference is sometimes made to the (positive) total of the negative flags. Unflagged metrics (and those not reportable) are assigned a flag value of zero.

z-score: defined as the ratio of the difference, d, (of a metric) to the standard deviation of the difference, std(d); the latter measure is derived according to a formula and also quoted in the outputs. The z-score is used to test whether the difference between a provider's observed indicator and its benchmark is statistically significant or not.

Annex A – Further discussion about investigating the feasibility of grouping providers by type

In Section 3.3.4, we discussed the comparability of TEF awards given that a wide diversity of providers (different provider 'types') are benchmarked together. In light of further evidence from Sections 10.1 and 10.2 suggesting different provider types may already be treated somewhat differently in the current TEF process, in the Conclusions (Section 11) we recommended that further research be undertaken on the feasibility and practicality of grouping providers by type. In this Annex, we provide some further thought on how any grouping might be achieved, and what should be considered in that work.

As guiding principles, we would see benefit in using groupings that:

- provide stability over time: providers that remain largely the same should be quite unlikely to change from one category to another at a future re-assessment, all other things being equal
- are based on absolute criteria as far as possible: the categorisation of one provider should be largely independent of that of others, or at least not determined by the specific subset of providers under consideration in the analysis (as an example, identifying similar providers from only those assessed in TEF Year Three would provide different groupings from an analysis considering all providers)
- are largely not based on self-selection of providers and not easily gamed by providers being able to change their characteristics slightly so as to move to a different group
- are meaningful and recognised as useful by TEF users (for example, potential students)
- are largely accepted by providers.

We consider two broad approaches that could be used to define provider types, whether separately or in combination:

- group providers by some typology
- group providers by data-driven measures.

The former would see providers grouped according to some categorisation, which may already be in existence. We don't have a deep understanding of existing classifications in the higher education sector but in this report we have already mentioned HESA's classification of higher education providers (higher education institutions (HEIs), further education colleges (FECs), and alternative providers (APs)). This simple classification provides an option for a grouping criterion that is objective, intuitively appealing, and may on its own result in immediate improvements in utility to users. However, we understand that use of that classification may end in the near future.

Data-driven approaches might be informed by techniques such as cluster analysis, which identifies providers that are similar to each other (according to some statistical measure) to form the same group (cluster), whereas providers in different clusters tend to be dissimilar to each other. A simpler data-driven approach could come through use of size measures, for example student numbers grouped into size-bands.

In terms of the desirable principles we set out previously, we consider use of an existing typology to be preferable in principle, though our knowledge doesn't include how any existing categorisation functions in practice, especially regarding self-selection and stability over time. On the possible data approaches, cluster analysis would be susceptible to producing groups that are not intuitive or are difficult to explain, and the clusters would likely change over time and be dependent upon the other

providers that were included in the analysis. The simpler approach of using size-bands appeals more in principle – it is easier to understand and likely more stable over time for most providers, but the choice of the size-band boundaries would need careful thought. Use of student numbers as the size measure could be beneficial though, as our analysis has showed various associations between it and TEF processes and outcomes. Perhaps a combination of approaches to give groupings such as Larger HEIs, Smaller HEIs, FECs, and APs would prove workable. If short-term movements between groups are likely because of small changes in providers' characteristics, then use of some 'frozen' version of the classifying variables may be beneficial with regards to stability

A more general consideration in grouping is statistical robustness: each group should contain enough providers (and students) to be considered robust, and be capable of producing TEF outcomes that are statistically significant in cases where an indicator truly differs from its benchmark (that is, provider metrics still being flagged, assuming flagging is to continue in TEF).

Our recommendation would be to consider and test various options for groupings, and re-running TEF processes at least as far as Step 1a to determine the likely impact.

Naturally, implementation of independent assessments within groups would lead to very different TEF outcomes. Providers' benchmarks would be re-set based on the new, smaller groups of providers, and that could, in theory, lead to a provider's old 'Step 1a Gold' being replaced by a new 'Step 1a Bronze' if that provider were better than average across all providers but worse than average within its new group. Issues like that would need careful handling, and it may be useful to re-consider the outcome classifications or rating scale(s) at the same time, along with other changes that are to be implemented in TEF based on our, or other, recommendations.

A new set of rules about when and how often any re-categorisation exercise takes place should also be specified. This should include instructions for handling providers that are new to TEF assessments and providers that substantially change their characteristics, for example as the result of a merger, or any other notable change in student size or 'type'. As with other aspects of TEF, transparency about the processes would be vital.

Our main conclusion on this matter is that further research is essential, and requires the expertise of those with a good knowledge and understanding of the higher education sector as well expertise and experience on the statistical side. Defining provider groups that have desirable characteristics and can be implemented is not an easy challenge, but one that has the potential to improve the comparability of TEF awards.

Annex B – Consideration of a more general framework and alternative approach to the binary nature of flag values

For each metric (by provider, and subject, if appropriate) the values of *z*, *d* and *E*, in conjunction with the metric's weight, are transformed into a flag value. It is helpful to consider the formula for this as either:

flag_value = *prop* x *weight*

or, more realistically but depending on context of use and definition of other functions

flag_value = sign(*d*) x *prop* x *weight,*

where *prop* = some function of *z*, *d* and *E*, and the weight is pre-defined (for example, 0.5, 1.0 or 2.0 and dependent upon the metric and method (2017 or 2018)).

The current approach can be considered a special case of this more general framework if *prop*(*z*, *d*, *E*) is defined as the step function given by:

$$\begin{aligned} prop0(z, d, E) &= 1 \quad \text{if } |z| \geq 1.96 \text{ and } \{|d| \geq 2.0 \text{ or } E > 0.97\} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

The use of a step function means that two providers with values of *z* or *d* that are almost identical but fall either side of the (arbitrary) +/-1.96 or +/-2.0 thresholds will receive all or nothing of the metric's weight. That's important, as it could (potentially) lead to the difference between a Gold and a Silver starting point (and not even the borderline Gold/Silver) derived in Step 1a, as seen in Section 3.7. This is undesirable, as statistical uncertainty means that very close values (such as *d* = 1.99 and *d* = 2.01, which are unlikely to be statistically significantly different nor considered materially different from each other) could lead to very different starting points in Step 1a.

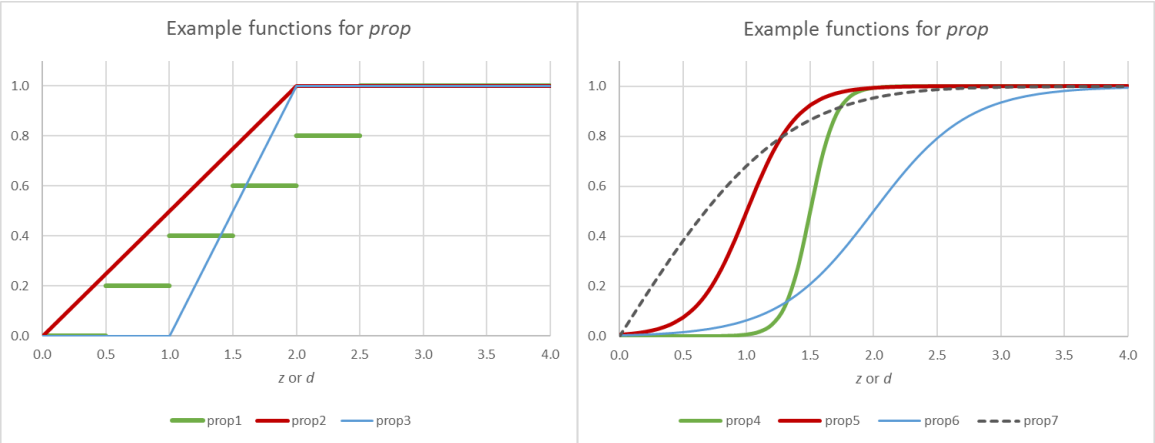
We have recommended in Section 4.1 consideration of alternative functions for *prop*. These should have more desirable properties, such as assigning increasing proportions (up to 1) of the weight to larger absolute values of *z* and *d*, and smaller proportions (approaching 0) for values of *z* or *d* near zero. It would likely assign the majority of the weight to values of *z* or *d* near the existing thresholds, and such a function could eliminate any requirement for defining thresholds for *z* and *d* altogether.

Desirable features for *prop* include:

- it is bounded between 0 and 1
- = 0 for |z| < some lower threshold, or approaches 0 as |z| decreases to 0
- = 1 for |z| > some upper threshold, or approaches 0 as |z| increases
- is a non-decreasing function.

We first consider examples of the function *prop* with only one input variable, which may be *z* or *d* (we note that *z* itself is a function of *d* (*z* = *d* / std(*d*)), but it may be more transparent to consider the two separately). The graphs in Figure 11 are shown for non-negative values of the input variable.

Figure 11. Illustration of seven example functions for *prop*, which all meet the listed desirable features but could result in different outcomes depending on their shape and location



prop1 is a more-refined step function than the current version.

prop2 and *prop3* are piecewise-continuous functions, each comprising all linear segments.

prop4, *prop5* and *prop6* are from the logistic family, (logistic being chosen for their sigmoid shapes), and tend towards 0 as *z* (or *d*) decreases (though don't attain 0 at z = 0, but can be made arbitrarily close) and tend towards 1 as *z* (or *d*) increases. As illustrated, these curves can be scaled and shifted as desired. Having a relatively flat and near-zero section for small $|d|$ would be desirable so that values of *d* that are not 'materially different from zero' do not attract much weight.

prop7 is defined as $2\,[\Phi(z)] - 1$, where Φ is the cumulative density function of the standard normal distribution. Use of such a function has a certain intuitive appeal, especially if used with *z* as the input variable, because of the link with standard normal probabilities. The 'doubling and subtracting 1' transformation of Φ is used frequently with normal probabilities to transform between one-sided and symmetric confidence intervals or hypothesis tests with normal distribution, and we note that a *z*-score of 1.65 yields *prop7* = 0.90; *z* = 1.96 yields *prop7* = 0.95 and *z* = 2.58 yields *prop7* = 0.99.

The need to incorporate both the variables *z* and *d* – as is used in the current methods – would require further consideration. One option would be to write the function *prop* as the product of two separate functions: *prop*(*d*, *z*) = *prop5*(*d*) x *prop7*(*z*), for example.

Many other options for functions of one or more variables are possible and warrant further thought and thorough testing. A further consideration would be whether the same function is provided for all metrics, or whether metric-specific functions should be applied.

Effects of *prop* functions on current metric calculations

If the function *prop* is defined such that the proportion for $|z| \geq 1.96$ and $|d| \geq 2.0$ is 1.0, then there will be no effect on values realised for those metrics that are currently flagged. But metrics currently scoring zero because $|z|$ was less than 1.96 or $|d| < 2.0$ would now receive some proportion (between 0 and 100%) of the full weight value. This can only leave unchanged or increase the total positive score of the metrics, and the likewise the total negative score. In turn, that will move providers' starting points towards the right and bottom of Figures 1 and 2 or, in general terms, towards the diagonal boundary in the diagrams.

If the function *prop* is defined such that some metrics with |z| ≥ 1.96 and |d| > 2.0 would now receive a value of less than 100% of the weight (*prop6*, for example), combined with those with |z| < 1.96 or |d| < 2.0 receiving a larger-than-zero value, the overall effect would be more mixed, with some realised flag values increasing and some decreasing. Indeed, it should be possible to include a scaling factor within the *prop* function so that the average provider metric remains the same as that currently observed, should that be a desired property.

If an approach such as use of the *prop* functions discussed were introduced, removing the current binary nature of the flags, it would change the intrinsic nature of a 'flag value'. As such, some alternative name to 'flag value' should probably be adopted even if the flags themselves are retained to help visually identify particularly good or weak metric outcomes.

Finally, use of a function, *prop*, if defined to have desirable mathematical properties, may lend itself to the approximation of standard errors for a combined indicator based on net total values, as discussed in Section 4.1. Having such measures of uncertainty included in the starting-point outcomes would be particularly useful.

Annex C – Distributions of reportable metrics for the 2018/19 subject-level TEF pilot

Table 22a. Distribution of number of reportable core metrics, 2018/19 subject-level TEF pilot

Subject	Providers	Percentage of providers with exactly *x* reportable metrics									
		x=0	*x*=1	*x*=2	*x*=3	*x*=4	*x*=5	*x*=6	*x*=7	*x*=8	*x*=9
Agriculture, food & related studies	57	0.0	0.0	1.8	1.8	1.8	1.8	8.8	10.5	19.3	54.4
Allied health	132	0.0	0.8	3.8	0.8	2.3	3.0	9.8	20.5	9.1	50.0
Architecture, building & planning	107	0.0	4.7	2.8	13.1	17.8	5.6	3.7	4.7	4.7	43.0
Biosciences	109	0.9	0.0	0.9	0.0	4.6	0.0	3.7	16.5	2.8	70.6
Business & management	225	0.4	1.8	3.6	3.6	5.8	1.8	7.6	11.1	7.6	56.9
Chemistry	50	2.0	0.0	0.0	4.0	4.0	0.0	2.0	6.0	6.0	76.0
Combined & general studies	20	0.0	0.0	15.0	15.0	20.0	5.0	5.0	10.0	0.0	30.0
Computing	170	0.0	0.6	2.4	1.2	8.8	2.9	4.7	7.1	8.2	64.1
Creative arts & design	175	0.6	1.1	1.1	4.0	2.9	1.7	3.4	7.4	9.7	68.0
Economics	69	0.0	2.9	0.0	1.4	5.8	0.0	4.3	8.7	5.8	71.0
Education & teaching	186	0.5	1.6	5.9	9.7	4.8	2.2	9.1	9.1	10.2	46.8
Engineering	184	0.5	2.2	8.2	7.6	25.5	1.1	2.7	3.8	3.8	44.6
English studies	94	0.0	0.0	0.0	1.1	5.3	1.1	0.0	4.3	1.1	87.2
General, applied & forensic sciences	38	0.0	5.3	0.0	2.6	7.9	2.6	0.0	15.8	10.5	55.3
Geography, earth, environment studies	67	0.0	0.0	0.0	3.0	0.0	1.5	1.5	3.0	4.5	86.6
Health & social care	126	0.0	3.2	1.6	0.0	3.2	2.4	7.1	12.7	19.0	50.8
History & archaeology	83	0.0	0.0	0.0	1.2	4.8	0.0	0.0	3.6	2.4	88.0
Languages & area studies	61	0.0	0.0	1.6	1.6	8.2	1.6	4.9	6.6	6.6	68.9
Law	99	0.0	0.0	0.0	1.0	4.0	1.0	2.0	7.1	3.0	81.8
Materials & technology	46	0.0	4.3	0.0	4.3	13.0	0.0	8.7	13.0	21.7	34.8
Mathematical sciences	66	0.0	1.5	0.0	0.0	6.1	1.5	1.5	9.1	3.0	77.3
Media, journalism, communications	100	0.0	1.0	0.0	6.0	1.0	1.0	4.0	10.0	6.0	71.0
Medical sciences	70	0.0	0.0	1.4	4.3	4.3	0.0	2.9	7.1	11.4	68.6
Medicine & dentistry	31	0.0	3.2	3.2	0.0	0.0	3.2	0.0	6.5	12.9	71.0
Nursing & midwifery	67	0.0	1.5	0.0	1.5	6.0	0.0	4.5	6.0	6.0	74.6
Performing arts	171	0.0	0.6	2.3	4.1	2.9	2.3	5.8	9.4	11.7	60.8
Pharmacology, toxicology, pharmacy	36	0.0	0.0	2.8	0.0	0.0	2.8	5.6	13.9	13.9	61.1
Philosophy & religious studies	74	0.0	1.4	0.0	4.1	8.1	2.7	10.8	17.6	8.1	47.3
Physics & astronomy	45	0.0	0.0	0.0	0.0	11.1	2.2	6.7	8.9	4.4	66.7
Politics	84	0.0	1.2	1.2	4.8	4.8	2.4	7.1	11.9	9.5	57.1
Psychology	109	0.0	0.9	0.0	1.8	0.9	0.0	4.6	6.4	3.7	81.7
Sociology, social policy, anthropology	117	0.0	0.9	1.7	2.6	1.7	0.9	4.3	15.4	8.5	64.1
Sport & exercise sciences	135	0.0	0.0	3.7	1.5	2.2	1.5	8.9	11.1	11.1	60.0
Veterinary sciences	20	0.0	0.0	0.0	0.0	0.0	5.0	15.0	15.0	15.0	50.0

Table 22b. Distribution of number of reportable split metrics, 2018/19 subject-level TEF pilot

Subject	Provider-splits	Percentage of provider-splits with exactly *x* reportable metrics									
		x=0	*x*=1	*x*=2	*x*=3	*x*=4	*x*=5	*x*=6	*x*=7	*x*=8	*x*=9
Agriculture, food & related studies	1,284	36.4	9.1	8.6	5.1	5.5	6.6	5.4	6.7	7.1	9.6
Allied health	2,940	31.3	9.1	8.6	4.6	5.2	8.9	7.1	7.9	5.3	11.9
Architecture, building & planning	2,345	37.6	10.4	9.6	4.6	3.6	5.6	5.7	6.4	5.9	10.4
Biosciences	2,542	23.1	5.9	7.6	5.2	4.6	8.2	6.3	11.4	8.5	19.4
Business & management	5,112	24.6	8.9	8.7	5.3	4.0	7.8	6.1	8.7	6.3	19.6
Chemistry	1,197	27.1	5.5	9.0	4.8	5.2	8.6	9.6	10.4	9.4	10.3
Combined & general studies	437	41.0	16.2	11.2	5.9	3.4	3.0	3.0	3.9	4.8	7.6
Computing	3,881	29.8	9.2	8.6	5.4	4.6	7.2	5.1	8.0	6.2	16.0
Creative arts & design	3,994	26.9	8.2	7.9	5.3	4.2	7.2	5.6	8.6	7.1	18.8
Economics	1,632	25.7	5.4	7.8	6.2	4.4	10.2	9.6	10.1	10.4	10.2
Education & teaching	4,075	36.9	11.2	10.4	5.3	4.4	4.8	5.0	6.3	4.0	11.6
Engineering	4,086	35.3	13.1	9.1	5.3	4.0	5.4	3.8	6.6	5.0	12.4
English studies	2,212	22.0	4.5	8.1	5.6	4.8	6.0	5.3	10.7	11.2	21.8
General, applied & forensic sciences	888	36.8	8.9	9.6	5.9	4.3	6.3	6.4	7.3	6.6	7.9
Geography, earth, environment studies	1,585	24.7	5.6	7.8	4.7	4.8	7.3	7.8	9.9	11.2	16.4
Health & social care	2,725	31.5	9.1	8.7	4.3	4.7	7.4	7.2	9.2	7.0	10.8
History & archaeology	1,949	25.6	3.8	8.3	4.5	5.2	6.0	6.3	10.3	10.2	19.7
Languages & area studies	1,437	29.9	7.9	9.5	5.7	6.5	6.1	5.9	7.9	8.7	11.9
Law	2,340	17.4	5.1	7.9	5.0	4.7	7.9	5.7	11.1	9.8	25.4
Materials & technology	1,068	41.0	9.1	7.7	7.2	5.5	5.4	6.6	6.7	4.6	6.1
Mathematical sciences	1,592	27.0	6.1	7.8	4.6	5.6	7.7	8.4	10.6	8.4	13.9
Media, journalism, communications	2,324	26.4	6.7	7.9	5.7	3.2	7.1	5.9	8.9	9.8	18.5
Medical sciences	1,641	24.6	7.6	9.0	4.5	4.3	7.5	10.0	11.1	8.7	12.7
Medicine & dentistry	731	17.4	3.8	5.6	4.1	2.2	12.2	11.6	14.9	17.5	10.7
Nursing & midwifery	1,546	20.4	4.7	7.2	4.7	3.2	5.9	5.6	10.9	9.6	27.9
Performing arts	3,903	29.4	7.4	9.8	5.8	3.9	7.3	6.4	9.5	7.1	13.3
Pharmacology, toxicology, pharmacy	855	30.8	6.3	6.9	4.9	3.9	8.7	9.1	8.8	9.9	10.8
Philosophy & religious studies	1,673	37.8	7.5	8.1	5.9	6.2	7.1	7.1	7.0	5.7	7.5
Physics & astronomy	1,064	33.3	7.5	7.8	5.8	4.5	7.5	8.4	8.7	9.1	7.3
Politics	1,927	27.3	7.3	8.8	5.8	4.8	9.2	7.0	9.1	9.8	10.9
Psychology	2,560	20.9	5.0	7.6	3.0	3.7	8.2	5.7	10.9	10.3	24.5
Sociology, social policy, anthropology	2,699	25.1	5.8	9.1	4.7	4.6	6.0	6.6	9.8	9.5	18.8
Sport & exercise sciences	3,013	29.8	10.6	8.4	5.3	4.6	6.7	5.9	8.2	7.0	13.4
Veterinary sciences	436	39.9	6.9	8.5	2.8	5.7	8.7	8.9	7.8	6.7	4.1

Annex D – Harmonisation. Comparison of Caring responsibilities, National identity, Religion and Sexual orientation with GSS standards

The Higher Education Statistics Agency (HESA) Student Record collects data on Caring responsibilities, National identity, Religion and Sexual orientation, but these topics are currently not included in the Teaching Excellence and Student Outcomes Framework (TEF), but potentially could at some future time. We review each of these topics here for completeness, comparing against Government Statistical Service (GSS) harmonised standards, but the recommendations we make do not form part of our evaluation of TEF as they do not currently feature there.

Section 6 of our report considers other topics that are currently featured in TEF.

Caring responsibilities

HESA uses a slightly different definition of caring to the definition recommended in the harmonised principles. The definition of a carer used by HESA is "The Carers Trust define a carer as: 'anyone who cares, unpaid, for a friend or family member, who due to illness, disability, a mental health problem or an addiction cannot cope without their support'."

This differs from the harmonised question in that the HESA definition includes caring for those suffering with addiction, and does not include caring for 'neighbours or others'.

Recommendation: Conduct research into the effect of including addiction and excluding caring for 'neighbours or others' in the definition of caring responsibilities, and consider amending question if this reveals significant discrepancies.

National Identity

The HESA question is broadly aligned with the harmonised question. No recommendations.

Religion

The questions used by HESA are broadly the same as the GSS harmonised questions appropriate to each UK country's Census, which will enable comparisons with other datasets specific to each country, but should note that the differences in wording of these questions will lead to comparability issues when comparing each Country against each other.

However, there are two differences between the response options used in the HESA questions and the GSS harmonised questions. Specifically:

- Christian – Baptist and Christian – Brethren, which have not been included in HESA's question
- Spiritual, which has been included in HESA's question.

These differences will not be substantive under the following two assumptions:

- those with Baptist and Brethren affiliations are captured in 'Christian – Other denomination' in HESA's question
- those that choose Spiritual in HESA's question are captured in 'Any other religion or belief' in data which uses the harmonised question wording.

Recommendation: Undertake further research into how the addition of the response option for 'Spiritual' will affect comparability with other sources.

Sexual orientation

There is a GSS harmonised principle for collecting data on sexual orientation by face-to-face or telephone interview. However, the 2021 England and Wales Census question differs slightly because it is optimised for self-completion and is based on more recent research, and therefore we recommend that this question design is used. The question used by HESA on sexual orientation differs from the recommended question in a number of ways:

There are differences in the wording of the question stems. It is recommended that surveys use the wording in the Census question ('Which of the following best describes your sexual orientation') as this more clearly acknowledges that the categories will not work exactly for all people's sexualities, but that they should choose the most appropriate.

The recommended question has 'Heterosexual / Straight' as a response option whereas HESA just refers to this as 'Heterosexual'. Census research found that the word 'heterosexual' is not familiar to everyone, and they have recommended reordering the wording to read 'Straight/heterosexual' as this aided understanding amongst the heterosexual population. They have worked with the LGB community to ensure this is considered acceptable.

The ordering of the response options is also different in the HESA specification, where it is alphabetical. In contrast, the recommended question and most other versions (for example, the NHS data dictionary) present the options broadly in order of population size (that is, heterosexual as the first option). It is common for ordering to cause data quality issues, and in this case some respondents will likely assume heterosexual to be the first option but actually select bisexual.

The response options in the HESA question separate out gay man and gay woman/lesbian – the harmonised version and the census have these as a single category. Given that sex is collected elsewhere, this distinction appears unnecessary.

It should be noted that the England and Wales Census will provide a better understanding of what is captured in the 'other' sexual orientation category, and this principle may subsequently be reviewed to consider the effects of incorporating more sexual orientations. In the meantime, it may be worth aligning with the census as lot of research has already been carried out on this, and will ensure comparability between the data sources.

Recommendation: With respect to harmonisation of statistical concepts:

- change the question wording to: 'Which of the following best describes your sexual orientation'
- reword 'Heterosexual' to 'Straight/heterosexual' to aid understanding of options
- order response options by size of population rather than alphabetically to limit respondent error
- do not separate out gay man and gay woman/lesbian, as data on gender/sex are gathered elsewhere.

Annex E – Example of mappings between JACS, HECoS and CAH classifications

To illustrate how the three subject classifications map between each other, we consider the area of mathematics as an example.

Under the Common Aggregation Hierarchy (CAH), mathematics is classified as:

- CAH Level 1: CAH09 Mathematical Sciences
- At CAH Level 2, there is only one nested group within CAH09:
    - CAH09-01 Mathematical Sciences.

    (Note: more generally, there can be more than one nested group)

- At CAH Level 3, there are three nested groups within CAH09-01:
    - CAH09-01-01 Mathematics
    - CAH09-01-02 Operational Research
    - CAH09-01-03 Statistics.

In terms of the Higher Education Classification of Subjects (HECoS), 15 subject codes currently map into CAH2 group CAH09-01:

100400	applied mathematics	101027	numerical analysis
100401	financial mathematics	101028	engineering and industrial mathematics
100402	mathematical modelling	101029	computational mathematics
100403	Mathematics	101030	applied statistics
100404	operational research	101031	medical statistics
100405	pure mathematics	101032	probability
100406	Statistics	101033	stochastic processes
		101034	statistical modelling

As noted in the Section 6.2.2, HECoS maps many-to-one in CAH, and no HECoS code splits into more than one CAH3 (and hence also CAH2) group. Thus, the 15 HECoS codes listed above are the only ones that map into CAH2 group CAH09-01, and none of those 15 codes maps anywhere else. Therefore, the set of those 15 HECoS codes and the single CAH2 group CAH09-01 may be considered equivalent. These principles seen in this example apply similarly to all HECoS and CAH codes.

For Joint Academic Coding System (JACS) codes, the mapping is not as neat. Those same 15 HECoS codes (equivalent to the one CAH2 group) comprise JACS codes: G000, G100, G110, G120, G121, G130, G140, G150, G160, G170, G190, G200, G290, G300, G310, G311, G320, G330, G340, G350, G390, G900, N300, all of which have mathematics-related names.

However, we note that these same JACS codes can also map into other HECoS codes (and hence into other CAH2 groups). In particular, and in relation to this example:

- G100 can also map to 100065 (Liberal arts)
- G121 can also map to 100430 (Mechanics), and
- N300 can also to 100107 (Finance) and 101040 (Risk management).

(Source: the mapping spreadsheets available at https://www.hesa.ac.uk/innovation/hecos)