Department
for Education

Teaching Excellence and Student Outcomes Framework: analysis of metrics flags

Research report

October 2017

Contents

List of figures

List of tables

Executive summary

The government has considered the lessons learned from Year Two of the Teaching Excellence and Student Outcomes Framework[1] (TEF), for which outcomes were published on 22 June 2017. The lessons learned exercise has focused on the practical operation of the TEF, in line with the commitments made earlier this year. The TEF Year Two lessons learned exercise focused on six main areas:

1. Whether the process of application and assessment worked smoothly and effectively;
2. Whether the guidance to providers was clear and understandable;
3. The way in which the metrics were used, in particular the use of significance flags and their role in generating initial hypotheses;
4. The balance of evidence between core metrics and additional evidence;
5. Whether commendations should be introduced for the next round of TEF assessments;
6. The number and names of the different ratings and their initial impact internationally.

The lesson learned report (https://www.gov.uk/government/publications/teaching-excellence-framework-lessons-learned) addresses all six areas and is supported by two key pieces of analytical work that address biases by provider and student characteristics, such as gender, ethnicity and age (https://www.gov.uk/government/publications/teaching-excellence-framework-analysis-of-final-award) and the weighting of the National Student Survey[2] (NSS) metrics in TEF (this report). Both feed into area three.

This research annex specifically looks at whether the significance flags from the National Student Survey (NSS) metrics have had a greater influence on the outcome of the final award than the other metrics. During the passage of the Higher Education Research Act (HERA), significant concern was expressed in Parliament about the use, or overuse, of the NSS metrics in the TEF. This concern has been echoed by significant numbers of sector stakeholders from all parts of the sector.

The analysis in this report uses a statistical methodology called Multiple Correspondence Analysis (MCA) to examine the influence that the significance flags from the NSS metrics have on the final award. This approach allows the contribution that significance flags make to overall patterns within the data to be identified. It also provides a graphical representation of categorical data showing differences between observed frequencies

[1] See http://www.hefce.ac.uk/lt/tef/whatistef/
[2] For more information see http://www.hefce.ac.uk/lt/nss/

and what is expected under the assumption of independence between metrics. This aids the detection and understanding of possible relationships between the significance flags and the other metrics. The analysis has been peer reviewed by Professor Gavin Shaddick, Chair of Data Science and Statistics at the University of Exeter. The peer review confirmed that the analysis was appropriate and robust, see Annex B for the full review.

The aim of the analysis is to identify metrics/significance flags that are highly associated with each other and to examine how much influence these have on the final year two TEF award. Where there is a strong association between two metrics there may be a case for reducing the weighting given to them in the TEF assessment in order to prevent them from having a disproportionate influence on the results.

Do the NSS metrics have a greater influence on the TEF final award than other metrics?

Key findings

- This analysis shows that there is a strong relationship between the core metrics of the TEF and the final award. There is some variation in the impact of individual metrics on the final award, with the National Student Survey (NSS) metrics having the strongest influence.
- The NSS metrics (‘teaching on my course’, ‘assessment and feedback’, ‘academic support’) have a large influence on the final award. We see that they are highly correlated i.e. being positive in one is associated with being positive in another. The bivariate correlations between the three NSS metrics are all greater than 0.95 whereas the correlation between the two employment metrics is markedly lower at 0.67.
- Providers are more likely to get a bronze award if they have negative flags for all three NSS metrics.
- Providers are more likely to get a gold award if they have positive flags for all three NSS metrics.
- Negative employment metrics are more influential on the final award than positive employment metrics.

Implications of the findings

1. In light of the findings we consider the weighting of the NSS metrics in TEF should be reduced. We consider that the impact it had on the final ratings was higher than would be desirable, considering the importance of the other metrics.

Furthermore, the very high degree of correlation means that in practice each NSS metric is not giving substantially new information compared to the other two NSS metrics.

We will therefore in future halve weighting of the NSS metrics, so that each NSS metric has a weighting of ½ whilst the other metrics have a weighting of 1.

2. In consequence, this means that the rules for calculating the initial hypothesis also need to be altered. The rules which we consider would create the greatest harmony with Year Two and that will therefore be applied are:

    a. For Gold, a provider would need overall metrics worth at least 2.5 (previously 3) and no negative metrics.

    b. For Bronze, a provider would need -1.5 or more negative metrics (previously -2).

    c. All other scenarios to receive an initial hypothesis of Silver (no change).

Introduction

The government has introduced the Teaching Excellence and Student Outcomes Framework (TEF) as a way of better informing students' choices about what and where to study for a higher education qualification. The TEF also aims to raise the esteem for teaching and recognise and reward excellent teaching. The TEF Year two final ratings were recently published[3] giving a rating of gold, silver, bronze or provisional to higher education providers that participated. Participation in TEF is voluntary and in total 295 providers participated. Of these, 231 applied for TEF assessment resulting in a gold, silver or bronze award and the remaining 64 providers opted for a provisional award as they did not have enough data for a TEF assessment.

The assessment process looks at core metrics, contextual information and additional information that is submitted by the provider. The data sets[4] used for the core metrics in Year Two of TEF are:

- the teaching on course (TEACH), assessment and feedback (ASSESS) and academic support (ACAD) scales from the National Student Survey (NSS). These will be referred to as the NSS metrics;
- retention using Higher Education Statistics Authority (HESA) UK Performance Indicators and the Individual Learner Record (ILR). This will be referred to as the non-continuation metric (NCON);
- proportion in employment or in further study (EMP) using Destination of Leavers from Higher Education (DLHE) survey. This (and the one below) will be referred to as the DLHE metrics;
- proportion in highly skilled employment or in further study (HSEMP) using Destination of Leavers from Higher Education (DLHE) survey.

Following the publication of the final year two results, it was agreed that a lessons learned exercise will be conducted to inform the implementation of TEF year three. This analysis looks at whether specific metrics have greater influence on the final award than others.

It should be noted that this analysis is only looking at the core metrics for each provider, The assessment process to arrive at the final award for a provider also includes metrics by subgroups such as ethnicity; deprivation; age and gender, contextual data and a submission from the provider. The submission allows the provider to add additional

[3] Results are published here http://www.hefce.ac.uk/tefoutcomes/
[4] Full details are set out in the TEF Year Two Specification: http://webarchive.nationalarchives.gov.uk/20170517113229/https://www.gov.uk/government/publications/teaching-excellence-framework-year-2-specification

context further to the standard contextual data, explain its performance against the core and subgroup metrics or further explore performance for specific student groups. The final award is not determined from the core metrics alone. The core metrics are used to form an initial rating for each provider which is then considered in light of the provider submission.

The aim of the analyses is to examine whether specific metrics and flags are associated with the final award.

The data and descriptive statistics

In year two of TEF, 295 providers participated and of these 231 had enough data to be awarded a full TEF award of gold, silver or bronze. In this analysis, the dataset consists of 231 higher education providers with a variable identifying the final award (gold, silver or bronze) for each provider and six metrics flag variables (three NSS, two DLHE and one non-continuation) indicating the flag (++, +, =, -, --, not reportable) for that metric for each provider, see Table 1. A metric has a non reportable flag if the provider does not have enough data for that metric. Each provider is given one of the six flag categories for each of the six metrics. Which flag a provider's metric gets is determined by comparing the provider's individual score (for a metric) with the benchmark, see the TEF year two specification[4] for full details. Benchmarks are used to allow meaningful comparisons between providers by taking into account the different mix of students at each provider. The benchmark is a weighted sector average where weightings are based on the characteristics of the students at the provider. See the HESA website[5] for full details.

The flag represents if the metric score is significantly and materially different from the benchmark. In TEF metrics, the number of standard deveiations that the indicator is from the benchmark is given as the z-score. A flag is allocated where the metric score is at least +/-2 percentage points from the benchmark and the z-score is at least +/- 1.96. A positive flag is labelled '+' and a negative flag is labelled '-'. Where the metric score is at least +/- 3 percentage points from the benchmark and the zscore is at least +/-3, the flags are labelled '++' or '--'. If the metric score is within +/-2 percentage points from the benchmark and the z-score is within +/-1.96, the flag is labelled '='.

Assessors use the flags to determine the initial TEF rating for each provider. A gold initial rating requires three or more positive flags and no negative flags. Providers with at least two negative flags are allocated bronze initially. In all other cases an initial rating of silver is allocated. This process gives equal weighting to each of the six metrics even though three of the metrics are from the same datasource (NSS).

Table 1 shows the number of providers for each metric and flag category.

[5] https://www.hesa.ac.uk/data-and-analysis/performance-indicators/benchmarks

Table 1 The six metrics (Academic support; Assessment and feedback; Teaching on my course; Employment; Highly skilled employment and Non-continuation) with flag categories for each metric.

Metric	Flag name	Number of providers with this flag	Definition
Academic support (NSS)	ACAD -	13	Provider has single negative flag for ACAD
	ACAD --	17	Provider has double negative flag for ACAD
	ACAD +	25	Provider has single positive flag for ACAD
	ACAD ++	31	Provider has double positive flag for ACAD
	ACAD =	141	Provider has a neutral flag for ACAD
	ACAD NR	4	Provider has a not reportable flag for ACAD
Assessment and feedback (NSS)	ASSESS -	14	Provider has single negative flag for ASSESS
	ASSESS --	21	Provider has double negative flag for ASSESS
	ASSESS +	24	Provider has single positive flag for ASSESS
	ASSESS ++	52	Provider has double positive flag for ASSESS
	ASSESS =	116	Provider has a neutral flag for ASSESS
	ASSESS NR	4	Provider has a not reportable flag for ASSESS
	TEACH -	15	Provider has single negative flag for TEACH

Metric	Flag name	Number of providers with this flag	Definition
Teaching on my course (NSS)	TEACH --	14	Provider has double negative flag for TEACH
	TEACH +	23	Provider has single positive flag for TEACH
	TEACH ++	20	Provider has double positive flag for TEACH
	TEACH =	155	Provider has a neutral flag for TEACH
	TEACH NR	4	Provider has a not reportable flag for TEACH
Employment (DLHE)	EMP -	9	Provider has single negative flag for EMP
	EMP --	11	Provider has double negative flag for EMP
	EMP +	28	Provider has single positive flag for EMP
	EMP ++	19	Provider has double positive flag for EMP
	EMP =	158	Provider has a neutral flag for EMP
	EMP NR	6	Provider has a not reportable flag for EMP
Highly skilled employment (DLHE)	HSEMP -	15	Provider has single negative flag for HSEMP
	HSEMP --	43	Provider has double negative flag for HSEMP
	HSEMP +	19	Provider has single positive flag for HSEMP

Metric	Flag name	Number of providers with this flag	Definition
	HSEMP ++	70	Provider has double positive flag for HSEMP
	HSEMP =	78	Provider has a neutral flag for HSEMP
	HSEMP NR	6	Provider has a not reportable flag for HSEMP
Non-continuation	NCON -	10	Provider has single negative flag for NCON
	NCON --	20	Provider has double negative flag for NCON
	NCON +	19	Provider has single positive flag for NCON
	NCON ++	15	Provider has double positive flag for NCON
	NCON =	161	Provider has a neutral flag for NCON
	NCON NR	6	Provider has a not reportable flag for NCON

Correlation between metric flag variables

Looking across provider scores, the correlation between the three NSS flag variables (using the counts from Table 1) is very high:

- academic support/assessment and feedback, correlation is 0.972
- academic support/teaching on my course, correlation is 0.996
- teaching on my course/assessment and feedback, correlation is 0.946

The correlation between the two DLHE metrics is markedly lower at 0.665. See Table 2 for correlations between all metrics.

Table 2 Correlation between metric flag variables

	ACAD	ASSESS	TEACH	EMP	HSEMP
ACAD	1	0.972	0.996	0.994	0.739
ASSESS		1	0.946	0.942	0.865
TEACH			1	0.998	0.678
EMP				1	0.665
HSEMP					1

Multiple Correspondence Analysis

A cross tabulation is a table of the frequencies within categories and combinations of categories that can help identify relationships between variables. It provides a method for checking whether the distribution of metric flags differs between providers. If there was no relationship between a provider and the distribution of the metric flags, i.e. they were independent, then the flags would occur with the same frequency for each provider. These are the expected values under the assumption of no relationship.

Multiple correspondence analysis (MCA) is a statistical technique for analysing the pattern of relationships between multiple categorical variables and, amongst other things, provides a graphical representation of cross tabulations. The overall goal of MCA is to understand differences between the observed data and what would be expected if there were no relationships between variables, that is, the expected values based on the assumption of independence. At the core of MCA is the production of new variables, known as dimensions. Each dimension is a combination of the original variables, e.g. the metric flags, and they are constructed so that the first explains as much of the difference between the observed and expected frequencies as possible. Subsequent dimensions are constructed, in order, to explain the maximum amount of the difference between the observed actual and expected values that remain (after the previous dimensions have been fit).

Method

Within MCA the dimensions are constructed using a frequency table known as a Burt table. A Burt table is a cross tabulation of each category (of each variable) against each of the others. A portion of the Burt table for the flags data is shown in Table 3. For example, it shows that 25 providers have a positive flag for academic support (ACAD) and 4 providers have a positive flag for ACAD and a positive flag for teaching on my course (TEACH). A metric is regarded as being not reportable if there are not enough students and in the MCA analysis the ‘not reportable’ (NR) category is omitted due to small numbers.

Table 3 Burt table showing a selection of cross tabulations of metrics and flags

		ACAD					…	TEACH					
	+	++	=	-	--	NR		+	++	=	-	--	NR
ACAD +	25	0	0	0	0	0		4	3	18	0	0	0
++	0	31	0	0	0	0		8	15	8	0	0	0
=	0	0	141	0	0	0	…	10	2	117	9	3	0
-	0	0	0	13	0	0		0	0	4	5	4	0
--	0	0	0	0	17	0		1	0	8	1	7	0
NR	0	0	0	0	0	4		0	0	0	0	0	4
⁞							⁞						
TEACH +	4	8	10	0	1	0		23	0	0	0	0	0
++	3	15	2	0	0	0		0	20	0	0	0	0
=	18	8	117	4	8	0	…	0	0	155	0	0	0
-	0	0	9	5	1	0		0	0	0	15	0	0
--	0	0	3	4	7	0		0	0	0	0	14	0
NR	0	0	0	0	0	4		0	0	0	0	0	4

Results

This section presents the results from the MCA analysis using the first five dimensions. Table 4 shows how much of the total difference between the observed frequencies and that what would be expected if there were no relationships between variables can be explained by the first five dimensions. The total number of dimensions is given by subtracting the number of variables from the total number of categories across all variables (30 – 5). Together, the first two dimensions explain over 60% of the difference with the first five dimensions explaining almost 90%. The remainder of the dimensions explain progressively less, each explaining only 1 or 2 % of the difference.

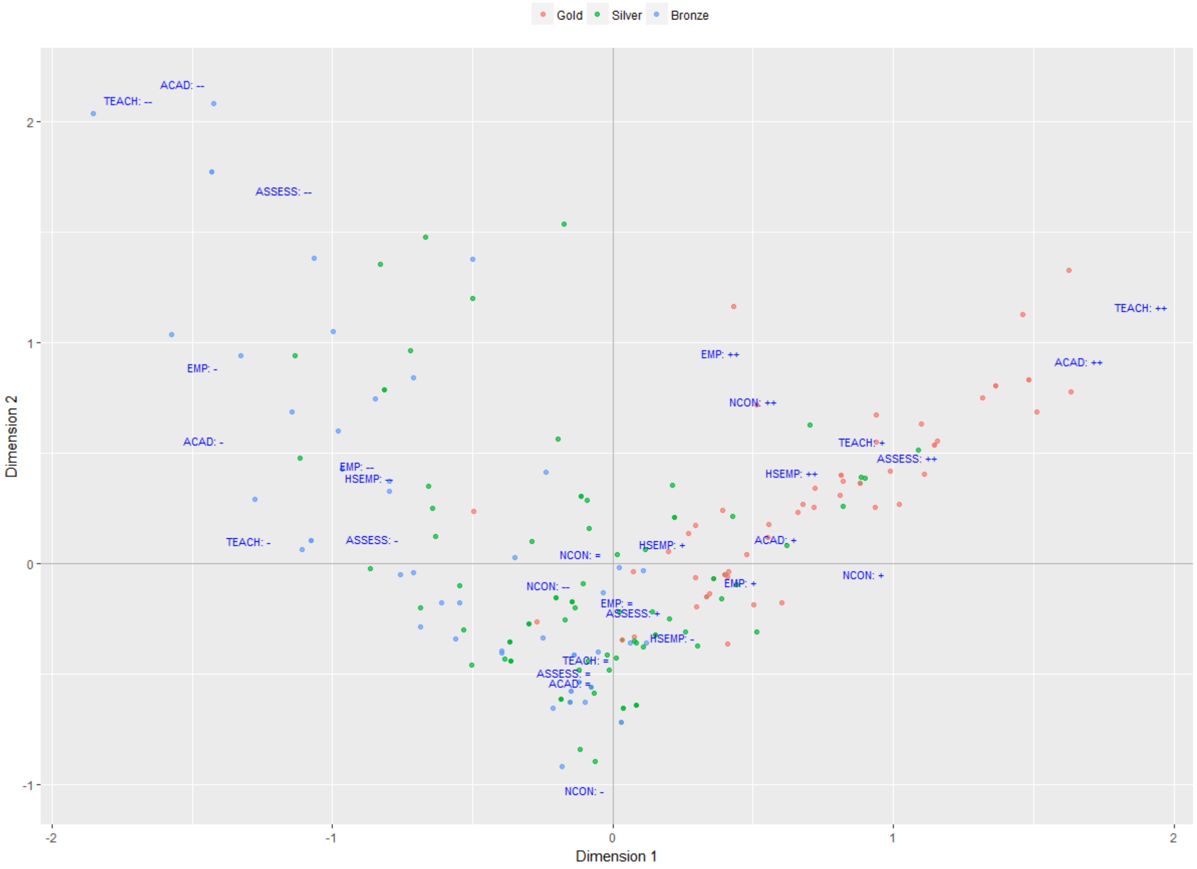
Table 4 Percentage of difference (between observed and expected frequencies) explained in each of the first five dimensions when the not reportable category is excluded. Also shown is the cumulative percentage difference.

	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
Percentage of difference	38.8% (38.8%)	23.9% (62.7%)	12.0% (74.7%)	8.0% (82.7%)	5.5% (88.2%)

Figure 1 shows the information contained in the first two dimensions in terms of the category of each metric, with the final award for each provider signified by coloured dots. The distances between points in this two-dimensional display show the extent to which the relative frequencies across rows and columns in the cross-tabulation are similar to each other. For example, the providers (dots) that are close together to the right of the plot have similar frequencies for metrics with a double positive flag, and providers that are closest together to the left of the plot have similar frequencies for metrics that have a double negative flag. Providers closest together in the centre of the plot have similar frequencies for metrics with a neutral, positive or negative flag.

Considering dimension 1, the flags that contribute most to distinguishing providers are the double positive flags for the NSS metrics, that is, the positive flags are to the right of the plot. The positive flags for the highly skilled employment metric (HSEMP) and non-continuation metric (NCON) also feature but do not contribute as much as the NSS metrics. The double negative flags (to the left of the plot) contribute the most to dimension 2.

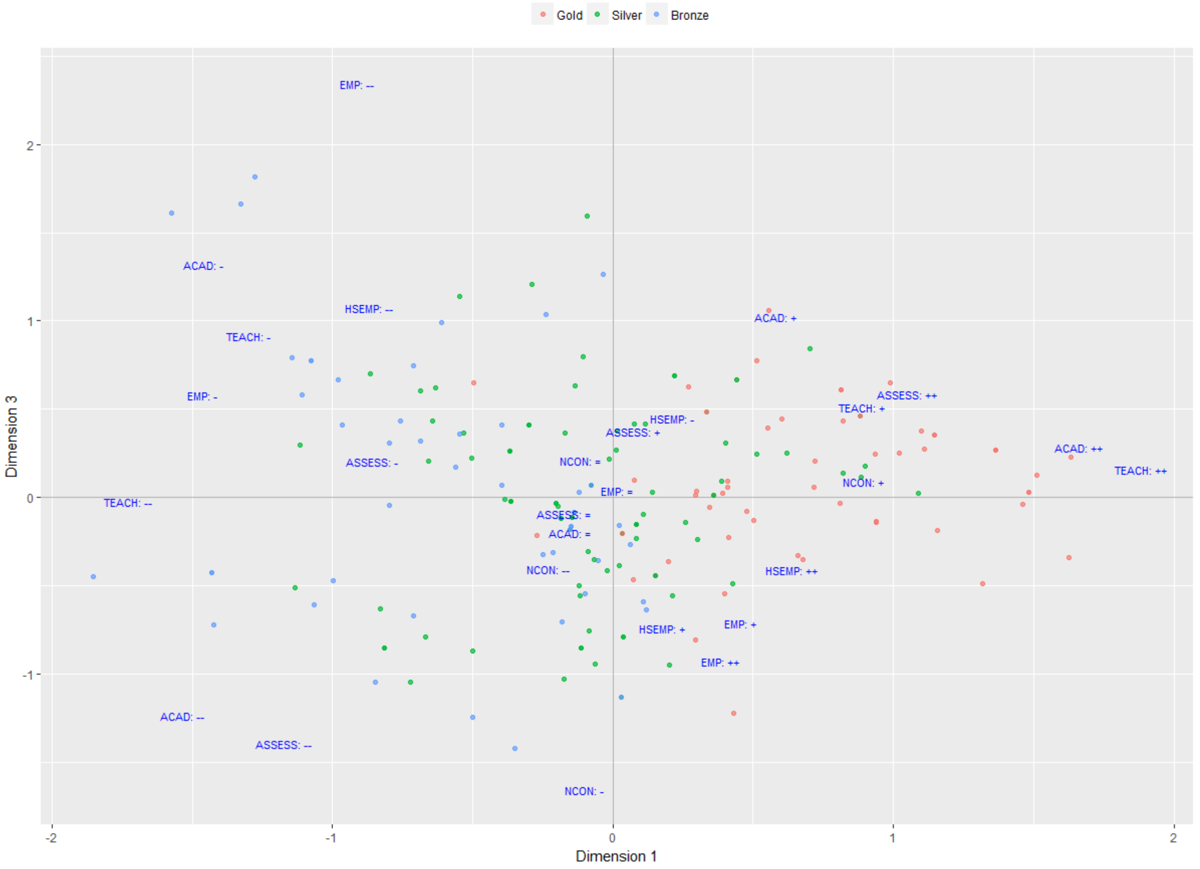
Figure 1 Map of metrics within dimensions one and two when the not reportable category is excluded. The coloured dots show the providers colour coded by the final award



Providers with a bronze award are clustered mainly around metrics with negative flags, whereas providers with a gold award are clustered around metrics with positive flags. Providers with a silver award are clustered around metrics with a mixture of neutral, negative and positive flags. **Within dimension 1 the NSS metrics with positive flags are on the edge of the geometric space (far right of the plot) meaning that they contribute more to distinguishing award types. Within dimension 2 it is the NSS metrics with negative flags that contribute more to distinguishing award types (top left hand side of the plot).**

Figure 2 shows the results for dimensions one and three. For dimension 3, the double negative flags for the DLHE metric EMP and the NSS metric ASSESS contribute most to discriminating providers. The negative flag for the non-continuation metric also plays a part but to a lesser degree. **Along dimension three, double negative flags for the DLHE metrics EMP/HSEMP and the non-continuation metric NCON discriminate providers the most**. See Table 5 for the contribution of categories to each dimension for the first five dimensions.

Figure 2 Map of metrics flags within dimensions one and three when the not reportable category is excluded. The coloured dots show the providers colour coded by the final award



Which metric and flags are most influential in each dimension?

Table 5 shows the contribution from each metric category in each dimension for the first five dimensions. The average contribution per category is 3.33 (100/30) so if each category had equal influence on a dimension then each category would have a contribution of 3.33. Here, just 3 (out of 30) categories, the double positive flags for the NSS metrics, contribute 39% to dimension one, i.e. they are highly influential in dimension one.

Table 5 Contributions of categories (excluding not reportable category) in each of the first five dimensions constructed to explain differences in the observed frequencies in the data and what would be expected if there were no relationships between variables. The larger the value, the bigger the contribution to that dimension. The largest values in each dimension are highlighted in bold

	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
ACAD: -	5.07	0.86	5.86	3.81	4.15
ACAD: --	7.31	**17.06**	6.73	0.49	6.80
ACAD: +	1.54	0.06	6.74	**15.02**	0.92
ACAD: ++	**15.65**	5.54	0.64	4.78	0.01
ACAD: =	0.60	8.69	1.50	0.16	0.39
ASSESS: -	1.88	0.04	0.15	0.09	1.37
ASSESS: --	5.29	**12.74**	**10.56**	0.03	4.34
ASSESS: +	0.02	0.25	0.85	**21.69**	0.70
ASSESS: ++	**10.48**	2.54	4.55	0.67	1.45
ASSESS: =	0.65	6.01	0.27	7.79	2.56
EMP: -	3.53	1.51	0.78	0.00	5.29
EMP: --	1.68	0.45	**15.63**	3.17	0.95
EMP: +	1.06	0.04	3.69	0.85	**12.77**
EMP: ++	0.51	3.67	4.24	2.36	4.86
EMP: =	0.01	1.01	0.06	1.40	**10.07**
HSEMP: -	0.12	0.35	0.78	0.20	6.44
HSEMP: --	5.95	1.38	**12.85**	0.25	2.04
HSEMP: +	0.11	0.03	2.71	3.50	0.60
HSEMP: ++	5.20	2.50	3.04	0.08	5.47

	Dimension 1	**Dimension 2**	**Dimension 3**	**Dimension 4**	**Dimension 5**
HSEMP: =	0.49	4.55	0.37	1.13	7.20
NCON: -	0.02	2.23	**7.09**	2.13	5.56
NCON: --	0.19	0.04	0.85	7.15	1.76
NCON: +	2.79	0.01	0.04	0.00	2.13
NCON: ++	0.69	1.73	2.15	8.91	2.05
NCON: =	0.37	0.07	1.81	0.18	0.77
TEACH: -	4.62	0.03	3.22	8.64	7.58
TEACH: --	7.64	**13.11**	0.00	0.07	0.03
TEACH: +	3.31	1.50	1.53	0.18	0.03
TEACH: ++	**12.96**	5.73	0.12	3.75	0.50
TEACH: =	0.26	6.25	1.16	1.53	1.21

Which metric (NSS, DLHE or non-continuation) is most influential in each dimension?

Table 6 sets out percentages of the contributions aggregated over the metric types (NSS/DLHE/NCON). In dimension one, the NSS metrics explain 77% of the differences between observed and expected frequencies (the latter calculated under the assumption of independence between the metrics) and the DLHE and non-continuation metrics explain 18% and 4% respectively. In dimension two, the respected contributions are 80%, 10% and 4%. This makes it clear that the NSS metrics are highly influential in most dimensions, but particularly that dimension 1 is based on positive NSS metrics (explaining 44% of the differences) and dimension 2 on negative NSS metrics (also explaining 44% of the differences). It is also worth noting that where the DLHE metrics are influential, it tends to be the below benchmark flags. There is high correlation (at least 0.95) between the three NSS metrics but the correlation between the two DLHE metrics is lower (0.67). So, we can conclude that in part, the over representation of the NSS metrics is due to a combination of their number (there are there 3 of them) and the lack of idependence between them, meaning that being positive in one means it is very likely that the others will be positive . In comparison, the employment metrics show a greater degree of independence.

To summarize, the NSS metrics contribute the most in terms of explaining patterns in the data, expressed as the differences between the observed and expected frequencies (the latter calculated under the assumption of independence between the metrics). The double positive NSS flags explain the largest portion of the total difference between actual and expected values with the double negative NSS flags explain the second largest. This means that the NSS metrics play a big role in allocation of an initial and final rating for each provider.

The DLHE metrics are not as influential as the NSS metrics but where they are influential it tends to be due to negative flags.

Table 6 Contributions of aggregated categories (sum of values in Table 5) in each of the first five dimensions constructed to explain differences in the observed frequencies in the data and what would be expected if there were no relationships between variables. The larger the value, the bigger the contribution to that dimension.

	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
NSS positive	44%	16%	14%	46%	4%
NSS negative	32%	44%	27%	13%	24%
NSS neutral	2%	21%	3%	9%	4%
All NSS	**77%**	**80%**	**44%**	**69%**	**32%**
DLHE positive	7%	6%	14%	7%	24%
DLHE negative	11%	4%	30%	4%	15%
DLHE Neutral	1%	6%	0%	3%	17%
All DLHE	**18%**	**10%**	**44%**	**10%**	**38%**
NCON positive	3%	2%	2%	9%	4%
NCON negative	0%	2%	8%	9%	7%
NCON neutral	0%	0%	2%	0%	1%
All NCON	**4%**	**4%**	**12%**	**18%**	**12%**

Annex A Technical definition of Multiple Correspondence Analysis

Techniques widely used to analyse categorical data are Chi-Square analysis, Fisher's exact test and the Ratio Test (or Z-test). However, the use of these statistical techniques depend on assumptions which which may be untenable and even if the assumptions are justifiable, the results of the analysis may be too general for interpretation. For example, a Chi-Square analysis (or any of the other tests) will indicate if there is a relationship between providers and the distribution of metric flags, but it will not indicate if one provider differs more than another provider. It also does not allow us to group the providers using the distribution of the metric flags.

The underlying concept of Multiple Correspondence Analysis the same as that of principal component analysis for use with categorical rather than quantitative variables. Using this technique relationships between row and column variables, together with relations between different levels of each variable, can be examined in a reduced dimensional space.

MCA is used to analyse a set of observations described by a set of categorical variables comprising several levels. MCA describes the patterns geometrically by locating each variable in a low dimensional space. It allows us to view the data using different angles (called dimensions). Categories that are closer in distribution are represented closer together in space. It is useful for uncovering groupings of variable categories in the dimensional spaces. The overall goal is to decompose the total difference between the observed values and expected values (i.e. the numerator of the Chi-squared statistic) by identifying a small number of dimensions in which the deviations from the expected values can be represented. Dimensions are "extracted" so as to maximize the distances between the row or column points, and successive dimensions (which are independent or orthogonal to each other) will "explain" less and less of the overall Chi-square value.

The number of dimensions created by MCA is equal to the number of dimensions in the original data but dimensions will contain differing amounts of information that can be used to explain the difference between observed and expected values. The dimensions that contained the greatest amount of information are the most important, the others can be ignored. Each dimension represents a combination of all the original variables but only some of the original variables will be strongly correlated with a particular dimension. The new dimensions allow the original data to be viewed from a different perspective thus enabling relationships between categories of different variables to be explored.

Annex B Peer review

The analysis in this report has been peer reviewed by Professor Gavin Shaddick, Chair of Data Science and Statistics at the University of Exeter.

Is the technical methodology appropriate for achieving the objectives of the research? For instance, are the proposed statistical techniques and model specification adequate? Is the analysis robust?

The choice statistical methodology used in this analysis is appropriate and reflects the nature of the data and the aims of the analyses. The analysis has been performed systematically and the interpretation of the complex outputs from performing dimension reduction has been made with care and with clear linkage to the overall aims.

Are the variables appropriate for achieving the objectives of the research?

The aim is to identify relationships between the allocation of awards and the different values of different metrics. The variables used in the analysis, apart from those not reportable for which there were small numbers, included the core metrics releated to the NSS and DLHE. Although the analyses could be expanded to include other variables, as the primary aim was to examine the relative influence of the NSS scores the variables included in the analysis appear entirely suitable.

Do you agree with the interpretation of the results, as set out in the report? Are the conclusions too strong or need further testing/revising?

The overall conclusions represent an accurate reflection of the results found in the analyses. The reporting of the highly complex outputs from the statistical modelling have been made with care and it is clear that a great deal of effort has been put into their interpration, explanation and in relating back to the overall aims.

Please include any other comments or suggestions, not covered above, here

One of the proposals that arises from these analyses is to reduce the weighting of the NSS metrics, and this seems justified given the high correlation between them. For future iterations of TEF, it may be useful to consider basing the choice of weights using a statistical approach. For example a model that predicts the probability that a provider be awarded gold might be used and the providers categorised according to their predicted probabilities of being in each group. Gold awards could then be allocated where providers had a probability of being in the gold category of at least 90%, or other suitable threshold. Senstivity analyses could be performed in order to assess the effects changes to the choices of weights, and cut-offs, in order to ensure that results were consistent with the aims of the exercise.

Department
for Education

Reference: DFE-RR736

ISBN: 978-1-78105-814-5

The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

Any enquiries regarding this publication should be sent to us at: www.education.gov.uk/contactus

This document is available for download at www.gov.uk/government/publications