Annex A: Differences in student outcomes: further characteristics

Data quality framework – a method for assessing the quality of student characteristic data

Enquiries to official.statistics@officeforstudents.org.uk

Publication date 4 June 2020

This annex should be read alongside the report 'Differences in student outcomes: further characteristics' (OfS 2020.30)

Contents

Summary **1**

Introduction **2**

Data quality framework **3**

Part I – Data availability 4

Part II – Data quality 6

Summary

Any feedback and queries related to this framework can be sent to William Rimington at official.statistics@officeforstudents.org.uk.

1. This annex provides detail of the framework developed to assess the quality and quantity of data related to the characteristics of students. The framework combines qualitative and quantitative methods of investigating data to aid decision making regarding its publication and further use.

2. This framework was developed as part of the report 'Differences in student outcomes: further characteristics'.[1]

[1] Available at: www.officeforstudents.org.uk/publications/differences-in-student-outcomes-further-characteristics/.

Introduction

3. The Office for Students (OfS) regularly publishes student data related to a set of protected characteristics (e.g. age, disability, ethnicity, sex) or characteristics that measure some kind of disadvantage (e.g. POLAR4[2]), including the differences in degree and employment outcomes.[3] This characteristic data plays a key role in our regulatory functions.

4. The characteristics mentioned above do not represent the entirety of the data held on students in higher education and individualised data on a number of other characteristics is also collected. Sources of this data include the Higher Education Statistics Agency (HESA) student record, HESA student alternative record (formerly AP record), Education and Skills Funding Agency Individualised Learner Record (ILR) and the Student Loans Company. As an additional source of information we can also link data from the Department of Education National Pupil Database (NPD) to data from HESA and the ILR.

5. Historically we have not published this data as part of outcome statistics due to data quality concerns. These concerns partly stem from the sector record being incomplete. Collection of data for some of these characteristics is not compulsory. Furthermore, for some characteristics, the ways questions have been asked in the past have not been consistent; this means there is uncertainty as to whether the responses are based on a consistent definition. In some cases, data will not be subject to data quality rules that are applied for certain characteristics and fields collected on individuals.[4]

6. To address these concerns we have developed the 'Data quality framework', which is a standardised method of investigating the quantity and quality of student characteristic data to aid in making decisions whether or not to publish data and whether to use it in further analyses.

7. This framework was developed by investigating characteristic data that is not currently used alongside commonly used student characteristics (disability, ethnicity and sex) to allow comparison to data that is deemed to be of sufficient quality for use and publication.

8. A large consideration in the data quality component of this framework is consistency of reporting. As such, this framework cannot be applied to new characteristics for which only one year of data exists. Caution should be applied when using newly collected characteristic data as applying this framework has shown that the first year of reporting is often of inconsistent quality.

9. Applying this framework represents one of the enhancements that the OfS is carrying out in order to improve its public sector equality duty (PSED) as outlined in the Equality Act 2010.

[2] POLAR4 is an area-based measure of young participation in higher education – see www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/about-the-data/.

[3] See www.officeforstudents.org.uk/data-and-analysis/differences-in-student-outcomes/.

[4] See https://www.hesa.ac.uk/collection/c18051/quality-rules.

Data quality framework

10. This framework can be used to help determine whether data for a student characteristic might be considered for use by the OfS. The framework is composed of two main parts:

    I. data availability

    II. data quality.

    As an overview, it can be applied by answering the following questions regarding the data.

Part I – Data availability

I.A – Documentation

- Is detailed documentation available for the characteristic?
- Is it clear how the data was collected and what student population it applies to?
- Are all the possible responses known and understood?

I.B – Disclosure rate

- For the eligible student population, has the characteristic been reported for the majority of students?

I.C – Provider responses

- Are all providers that could report data for the characteristic doing so?
- If not, is there a reason for this absence of reporting?

Part II – Data quality

II.A – Identified data issues

- Are there any issues associated with the data that would caution against its use for certain or all academic years?

II.B – Reporting consistency

- Year on year, how consistent are the proportions reported by providers for each of the categories of the characteristic?
- If there are inconsistencies, can they be explained?

II.C – Comparison to general population

- How do the sector proportions compare to the general population?
- Are any differences or similarities understandable given what we know of the student population?

11. The framework combines both qualitative and quantitative analyses. Some parts require the calculation of statistics whereas others require an understanding of the data and how it was collected. This being the case, the overall outcomes of applying the framework are open to interpretation and whether the data is deemed of sufficient quality will be dependent on its intended use.

Part I – Data availability

I.A – Documentation

12. The characteristic data should be properly documented with detailed knowledge of how the data was collected.

    a. **Collection population** – It is essential to know who the characteristic applies to as this will impact disclosure rates: a provider that exclusively teaches postgraduate qualifications would not report data for a characteristic that is only collected for undergraduate students so should not be included when calculating disclosure rates.

    b. **Collection history** – Each year the collection of data may differ with possible changes in categories and the population for which the data is collected. Researching the collection history of the characteristic is important for determining whether time series data can be used and whether differences year on year are the result of genuine trends or changes in data collection.

    c. **Collection policy** – It is important to determine whether collection of the data is compulsory or optional and how this data is collected. Some characteristics are collected directly by the providers and others via the Universities and Colleges Admissions Service (UCAS). This knowledge can be useful when investigating which providers do not report data and why, for example if a provider does not use UCAS they are unlikely to report data for a characteristic primarily collected via UCAS.

    d. **Question** – It is useful to know how the question regarding the characteristic is asked and whether the same question is asked by all providers or whether providers can choose how they collect the data.

    e. **Categories** – The possible categories for the characteristic must be known and understood.

    f. Investigating the documentation associated with the characteristic will likely provide useful information for Part II.A of the framework (identified data issues).

I.B – Disclosure rates

13. Data should be available for the majority of students.

    a. Disclosure rates represent the proportion of eligible students for which the characteristic data exists. They are calculated as follows:

$$100 \times \frac{\text{number of students for which characteristic is known}}{\text{number of eligible students for which the characteristic could be known}} = 100 \times \frac{\text{Knowns}}{\text{Knowns + Unknowns}}$$

    b. Depending on the intended use for the data, a decision should be made at this point regarding how to treat active responses that do not provide information (e.g. Information refused, Don't know). These could be classified as a known response as they should result from a student actively saying they do not wish to answer. Alternatively, as these responses do not provide information on trends and can impact calculations of proportions they could be treated as unknown.

    c. Where data is missing or the category represents missing data (e.g. No response given) then the data should be classified as unknown.

    d. The disclosure rate does not need to be 100 per cent and this is unrealistic for non-compulsory fields. However, a response should be recorded for the majority of students.

    e. A set disclosure rate cut-off has not been determined as part of this framework – an appropriate proportion for responses is dependent on the intended use of the data.

    f. Additionally, a disclosure rate cut-off can be biased by provider size. Large providers not reporting data will have a greater impact on disclosure rates than small providers not reporting data.

    g. This being said, the utility of the data with a disclosure rate below 67 per cent should certainly be questioned as this indicates that the characteristic is unknown for over a third of eligible students.

    h. It is also useful to calculate the disclosure rate for only the providers that report data in addition to all the providers that could be reporting data. If the characteristic has a low disclosure rate for the providers that reported data then this indicates that there may be anomalies in the reporting of data, for example providers reporting data for only a selection of their eligible students.

I.C – Provider responses

14. All eligible providers should report data. If a provider does not report data we should be able to determine why they do not.

    a. Not all providers will report data for a characteristic – a particular characteristic may not apply to their students. Alternatively, if reporting is optional they may choose not to collect/report data. For the characteristic data to be considered representative at a sector level it should be reported by the vast majority of providers that it is applicable to. If the data is not available for all providers, for this criterion to be satisfied, there should be a reason for the absence. For example, if a characteristic is primarily recorded as part of the UCAS

application, and the characteristic data is available for all providers except those that do not use UCAS, then this absence of data for those providers can be rationalised.

b. The impact of providers not reporting data on the sector-level statistics is unknown. This is because there is no way of knowing the makeup of their study body. For example, a provider with low numbers of – or poor outcomes for – underrepresented students may not report data, meaning the sector-level statistics calculated would be higher than in reality. Should data for a characteristic be determined to be of high enough quality for use then it is advisable that collection of that characteristic is made compulsory. Not only would this improve the disclosure rates but would also mitigate against providers choosing to not report data for a characteristic which may reflect badly on them.

c. Note this part of the framework only applies to data reported by providers and does not need to be considered when linking data from other sources, for example the NPD. However, school type should be considered when creating the eligible student population for NPD characteristics. For example, free school meals data is rarely reported by independent schools so should not be considered when calculating disclosure rates for this variable.

Part II – Data quality

II.A – Identified data issues

15. The data should be high quality.

a. Through investigation of the documentation (Part I.A) and observation of the proportions reported by individual providers, obvious issues/concerns regarding the data quality will be uncovered. While this is a subjective criterion, it is useful for excluding data entirely for years when it is obvious that it is of low quality. Where concerns are detected, efforts should be made to investigate these to decide whether they are serious enough to exclude use of the data.

b. These data issues should be evidence-based, resulting from investigation of the data and associated documentation. They should not be based on any pre-existing understanding or anecdotal evidence regarding the data quality.

c. It can also be useful to consider issues that may exist regarding how the questions are answered by students and how this may influence the data. Students may decide to answer questions in an inconsistent way, which may lead to under- or over-reporting for categories within a characteristic.

II.B – Reporting consistency

16. The data reported by providers for a characteristic should be relatively consistent as student populations are unlikely to change substantially year on year at a provider.

a. While it is reasonable to anticipate that, for a given characteristic, the proportions reported by a provider will change slightly, for the most part the proportions for a given year will likely be similar to the year before. If all providers are reporting data consistently (i.e. there are only small changes year on year for each category) then the sector-wide proportions will also change little year on year. If the opposite is true and the proportions reported by a

provider vary considerably year on year then the quality of the data should be questioned. For example, historically it has been possible for providers to report that 100 per cent of their entrants were the same gender as at birth one year, to the following year reporting that 1 per cent of their entrants were the same gender as at birth. If the variability of proportions is small we can be more confident regarding the quality of the data.

b. An inconsistency score can be produced by calculating the weighted standard deviation of differences in reporting between one academic year and the next. This is done by finding the difference in the proportion reported by a provider for a category of a characteristic between one academic year and the next. By calculating the standard deviation of all the differences for all providers, weighted by provider size, the variability of reporting for a characteristic can be examined. The equation for this calculation can be found below and the process is illustrated in Table A1.

$$Inconsistency\ score = \sqrt{\frac{\sum w_i (x_i - \bar{x}_w)^2}{\sum_i w_i}}$$

Where:

x_i = The difference between the proportion of students at a provider belonging to a characteristic category between one academic year and the next (e.g. if a provider reports 1% of their entrants are care experienced one year and 3% the following year then $x_i = 2$).

$\bar{x}_w$ = The weighted average difference in proportions reported for a category by all providers between one academic year and the next.

w_i = The number of students at the provider in the earlier of the two years being compared. This number is the sum of the students belonging to all the different categories for the characteristic but only includes those students for which the characteristic is known.

c. This method looks at the differences in reported proportions by a provider. As such differences in the student populations between providers do not influence the calculation (see Table A1).

d. The score is weighted to account for the size of the provider. A small provider is more likely to experience larger changes in their proportions for a category each year compared to a larger provider. Thus, without weighting, smaller providers would have an abnormally large impact on the sector level consistency score.

e. Weighting also accounts for any providers only reporting data for a very small proportion of their students (e.g. in the first year of reporting care experience data, some providers only reported data for their care experienced students and put their non-care experienced students as unknown, leading to provider proportions suggesting that 100 per cent of their students of known care status were in care).

f. The number of students in the earlier of the two comparing years is used for weighting to allow comparison of a new year of data to the proportions and amount of data from the year before. This allows us to account for the situations where providers report abnormally low amounts of data one year, leading to abnormal proportions. Weighting by the earlier of the

comparison years prevents any differences from the abnormal proportions being given a higher weight, which would occur if the later of the two years were used for weighting. For example, if a provider is reporting abnormal statistics because they report data for a very small number of their students, but the next year they report data for all their eligible students, the large differences will be given a small weight based on the earlier year.

g. If a provider stops reporting data (for example if they close) they are not included in the calculation so that this will not impact the inconsistency score.

h. When a characteristic is binary (only two categories, for example whether or not the student received free school meals), the consistency score will be identical for either category and for the characteristic as a whole. This is because when a proportion for a category increases, the other category will decrease by the same amount, and vice versa. Furthermore, the score is weighted by the same number of students regardless of the size of the category.

i. When investigating characteristics with multiple categories, it should be considered that inconsistent reporting for a category may be masked in the inconsistency score for the whole characteristic, if the change in reporting is spread across multiple categories and not a single category. In this scenario a category could be inconsistently reported but the characteristic inconsistency score would be low. For this reason, when considering multiple category characteristics it is important to create inconsistency scores at both the characteristic and category level. These can then be compared to determine whether it is likely that the above scenario is occurring. If a category has a large inconsistency score but the characteristic has a low overall score then that category is changing more than the other categories. Category level scores should be similar to the characteristic level score and the range of inconsistency scores at the category level should be small if all the categories are changing in a similar way. For a uniformly changing characteristic with many categories, the characteristic level score will be smaller than the range of the category scores and will be towards the median category score.

j. A category having an inconsistency score higher than the other categories does not necessarily mean the data is low quality; it can be correct. For example, when looking at ethnicity, White students have a higher inconsistency score than Asian, Black, Mixed and Other. This is as expected, as access to higher education for minority ethnic groups is increasing which means the White ethnicity group is changing to a greater extent than Asian, Black, Mixed and Other.

k. Calculating the inconsistency score does not detect individual providers reporting large differences year on year but summarises the characteristic on the sector level, as the framework is interested in whether data for a characteristic is useable as a whole. If a small number of providers are reporting large differences then this will not be detected, but if it is a common theme of the characteristic then this method will allow detection.

l. This method of detecting inconsistent data reporting allows for changes in the student population as a result of access and participation plans leading to a change in the study

population.[5] These changes are expected to occur over time and will not result from a single change between one year and the next.

m. Like disclosure rates, a set consistency score cut-off has not be set as this is dependent on the circumstances of the data in addition to the planned use. For example, the data on ethnicity, which is deemed to be high quality and regularly used in our regulatory functions, varies in score between 1.0 and 2.8.

n. As with Part IC of the framework, this part of the framework is not directly applicable when using NPD data as the data is not reported by providers but comes from linking data. It is useful to calculate the inconsistency score, but a high score does not necessarily mean the data quality should be questioned in the way it would for a provider reported characteristic.

Table A1: Examples of calculating inconsistency scores for a single category of differently reported characteristics

		2015-16	2016-17	Difference (x_i)	Number of students at provider in 2016-17 (w_i)
Consistently reported characteristic – similar across providers – good quality					
Provider A	Category 1	50%	53%	3	1,000
Provider B	Category 1	47%	48%	1	200
Provider C	Category 1	49%	49%	1	500
					Inconsistency score = 0.98
Consistently reported characteristic – variable across providers – good quality					
Provider A	Category 1	95%	93%	2	1,000
Provider B	Category 1	3%	4%	1	500
Provider C	Category 1	50%	49%	1	200
					Inconsistency score = 0.49
Inconsistently reported characteristic – questionable quality					
Provider A	Category 1	98%	1%	98	1,000
Provider B	Category 1	1%	97%	96	1,000
Provider C	Category 1	1%	1%	0	500
					Inconsistency score = 38.8
Characteristic consistently reported at large providers but not at a small one – weighting stops small provider skewing data					
Provider A	Category 1	70%	71%	1	10,000
Provider B	Category 1	73%	72%	1	10,000
Provider C	Category 1	75%	0%	75	5
					Inconsistency score = 1.17

[5] For more information, see www.officeforstudents.org.uk/publications/transforming-opportunity-in-higher-education/.

II.C – Comparison to general population

17. Where possible, comparisons should be made between the student population and the general population.

    a. The differences or similarities found should be considered and reasoned whether they are as we would anticipate, given our pre-existing knowledge of the student population. This comparison should not be overly detailed and for the most part should be used as a sense check.

    b. When making comparisons it is not necessarily the case that the proportions for the student population and general population should be the same. For example, a larger proportion of the student population is from socio-economic classification 1 (Higher managerial and professional occupations) compared to the general population. However, given that socio-economic background has a large impact on access to higher education, this is understandable. If these proportions were the same then the quality of the data would be questioned.

    c. An additional consideration when making comparisons is that some efforts to target underrepresented groups could ultimately lead to increased proportions for that group when compared to the general population.

    d. If needed, the student population should be limited to match the general population statistics that are available.

    e. This part of the framework is not essential and does not need to be performed if relevant comparison statistics cannot be found.