Chapter 6
Data quality

This chapter examines issues surrounding the quality of the OECD's international education data. It begins with a declaration of the OECD commitment to data quality and the quality framework used to collect, compile and disseminate education data. It then discusses the type of data-quality problems that arise and why they arise, and describes how the OECD assesses and addresses these issues. The chapter includes suggestions about making estimates for missing data and concludes with an account of the main data-quality issues that remain to be tackled.

This chapter examines issues surrounding the quality of the OECD's international education data. It begins with a declaration on the OECD commitment to data quality and the quality framework used to collect, compile and disseminate education data. It then discusses the type of data-quality problems that arise and why, together with a description of what the OECD does to assess and address them. It then makes some suggestions about making estimates for missing data and concludes with an account of the main data-quality issues that remain to be tackled in the area of international education data.

6.1 OECD dimensions of data quality

Data quality is fundamental to the credibility of the statistics produced by the OECD in general and by the OECD Directorate of Education and Skills in particular. The OECD collection of education statistics adheres to the core values stated in the OECD's *Quality Framework and Guidelines for OECD Statistical Activities* (OECD, 2011).

The OECD's education statistics are compiled and made available on an impartial basis. They are produced according to strictly professional considerations, including scientific principles and professional ethics, with regard to methods and procedures used for the collection, processing, storage and dissemination of statistical data.

Quality is defined as "fitness for use" for users' needs. This definition is broader than has been customary in the past, when quality was equated with accuracy. It is now generally recognised that there are other important dimensions. Even if data are accurate, they cannot be said to be of good quality if they are produced too late to be useful, or cannot be easily accessed, or appear to conflict with other data. Thus, quality can be seen as a multi-faceted concept. Which quality characteristics are most important depend on users' perspectives, needs and priorities, which vary across groups of users.

The OECD views quality in terms of seven dimensions: relevance, accuracy, credibility, timeliness, accessibility, interpretability and coherence. Last but not least, cost-efficiency is an important factor although not strictly speaking, a quality dimension. Cost-efficiency must be considered in the possible application of any one or more of these seven dimensions.

The OECD *Quality Framework* is therefore built around eight considerations:

- **Relevance:** measuring relevance requires the identification of user groups and their needs.
- **Accuracy** is the degree to which the data correctly estimate or describe the quantities or characteristics that they are designed to measure.
- **Credibility** is the confidence that users place in data products based simply on their image of the data producer, i.e. the brand image. Credibility is determined in part by the integrity of the production process. Principle 2 of the UN Principles of Official Statistics (UNESCO, 1994) states: "to retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data".
- **Timeliness** reflects the length of time between data becoming available and the events or phenomena they describe. The notion of timeliness is assessed on the time period that permits the information to be of value and still acted upon.
- **Accessibility** reflects how readily data products can be located and accessed from within OECD data holdings.
- **Interpretability** reflects the ease with which users may understand and properly use and analyse the data. The adequacy of the definitions of concepts, target populations, variables and terminology underlying the data, and information describing the limitations of the data, if any, largely determines the degree of interpretability.
- **Coherence** reflects the degree to which the data are logically connected and mutually consistent.
  - **Coherence within a dataset** implies that the elementary data items are based on compatible concepts, definitions and classifications and can be meaningfully combined. Incoherence within a dataset occurs, for example, when two sides of an implied balancing statement, such as inflows and outflows, do not balance.

  - **Coherence across datasets** implies that the data are based on common concepts, definitions and classifications, or that any differences are explained and can be allowed for.
  - **Coherence over time** implies that the data are based on common concepts, definitions and methodology over time, or that any differences are explained and can be allowed for. Incoherence over time refers to breaks in a series resulting from changes in concepts, definitions or methodologies.
  - **Coherence across countries** implies that the data are based on common concepts, definitions, classifications and methodology, or that any differences are explained and can be allowed for.
- **Cost-efficiency** measures the costs and provider burden relative to the output. Provider burden is a cost that happens to be borne by the provider, but is a cost nevertheless. Although the OECD does not regard cost-efficiency as a dimension of quality, it is a factor that must be taken into account in any analysis of quality as it can affect quality in all dimensions.

6.2 Types (or causes) of data-quality issues

As with any data collected by the OECD, the quality of education statistics and indicators disseminated depends on two aspects: the quality of the national statistics received and the quality of the internal processes for the collection, processing, analysis and dissemination of data and metadata. While the latter is within OECD's control, the former is less so.

The quality of national statistics received will essentially be a function of:

- the adequacy of national data sources to provide the required international data
- the extent to which international data definitions and guidelines are correctly applied
- internal capacity within countries to implement OECD guidelines and develop appropriate data collection systems
- the quality and reliability of data transfer channels between national statistics offices and the OECD.

Within the field of education, a number of factors may mean national data sources are inadequate to provide the required data at the international level.

- **The coverage of national sources** – either individually or collectively – may not match the intended coverage of education as defined in Chapter 3. This can result either in gaps in the reported data or over-reporting through the inclusion of educational programmes that are not in scope of the data collection. This can also happen where there is ambiguity surrounding the validity for inclusion of some programmes, such as some continuing education programmes. As countries will typically use a number of national data sources to compile their international data returns, inconsistent coverage between them can cause problems of internal consistency and perhaps double counting of data reported by an individual country. This may occur between student data at different ISCED levels or between enrolment and finance data.
- Similarly, **the point in time when the data are collected** (the reference period) and the date on which the count of students is taken (the reference year for statistics), may differ from the international requirements. Data may simply not yet be available for the intended reference periods of the data collections, either because the national data-processing timetable does not fit well with the international data collection or perhaps because national data collections do not occur every year.
- **National data item definitions** (e.g. of teachers, graduates and programmes) and **their classifications** (e.g. programme level or type of educational personnel) may be different from international guidelines.

Difficulties adhering to international guidelines can arise when national data cannot readily be translated into the international definitions, but they can also arise from weaknesses in the guidance itself. This may be due to the lack of an internationally agreed definition for a data item or a lack of clarity in its description.

In addition to these challenges, ensuring education statistics are comparable over time is often a challenge. There are three possible reasons for significant changes in the data from one year to another:

- **Changes in the educational system**. This refers to "real" changes in the data due to changing conditions of the educational system, such as the implementation of reforms that lead to, for instance, an increase in the stock of students.

- **Changes in the coverage of the data collection**. This refers to changes introduced due to the exclusion or inclusion of programmes compared to the previous year, for example the inclusion of adult literacy programmes or private schools.
- **Changes in the methodology used**. This refers to significant changes in the data due to new/modified methodologies in collecting or estimating data.

6.3 Tackling data-quality issues

Both the OECD and member countries have committed considerable efforts to assuring and improving the quality of education data. On the one hand, this involves a rigorous data collection and verification process and on the other hand, a commitment to continuously address areas of weakness in data quality.

The OECD's main actions to improve data quality are:

- Meeting with countries to provide advice and guidance on detailed data definitions and data reporting. This guidance advises the data providers about the checks that will be carried out and the treatment of missing values.
- Using electronic data collection instruments (electronic questionnaires) which include aggregations of sub-classifications in areas where it is known to be difficult for countries to provide the required data, for example the disaggregation of some ISCED levels. Those instruments allow checks to be readily available. It also helps with coherence across the different questionnaires. For instance, student enrolment data are collected on different bases to match the coverage of the finance and the personnel data.
- Using codes in data tables to inform users of missing data or data of lower quality:
  - category not applicable (a)
  - data included in other categories (x, xr…, xc…, xa… indicating the row (r) and column (c) in which the data are included)
  - includes data from another category (d)
  - data not available (m)
  - too few observations to provide reliable estimates (c)
  - values are below a certain reliability threshold and should be interpreted with caution (r)
- Asking countries to provide metadata along with their data which outline the concepts, definitions and methods used in collection, compilation, transformation, revision practices and dissemination of statistics. For education statistics, an important element of metadata is the mapping of countries' national educational programmes to the ISCED levels and the description of these programmes. Other metadata collected include:
  - reference periods (start and end of school years) for each level of education
  - data collection periods (e.g. snapshot or whole year counts within the reference periods)
  - reference data for student ages
  - theoretical starting, ending and graduation rates
  - data sources and methods used
  - documentation on breaks in time series.
- Including automated verification in the electronic questionnaire spreadsheets sent to countries to fill in. The data providers can then run a check routine which identifies data cells with missing values and verifies the internal consistency of the data both within and between tables. Countries are asked to explain any verification errors remaining in their questionnaire submission.
- Subjecting the submitted questionnaires to rigorous scrutiny from the OECD Secretariat, particularly checking year-on-year consistency of the data, and raising queries with countries as required. These may lead to countries resubmitting data.

- Informing countries on how their data have been used in the calculation of the indicators that will subsequently appear in the publication *Education at a Glance* through preliminary tables shared with countries. Countries' knowledge of the use to which the data will be put is an important element in achieving good data quality.

Beyond the data collection process, the OECD makes a continual effort to assess and to improve the data quality, mainly conducted through the agendas of the INES Working Party and INES Network meetings. Special studies are conducted in areas where comparability problems have been identified. This specific approach allows the OECD to clarify countries' current data reporting approaches and use this to refine the data reporting guidance it provides to countries and to enrich the metadata. Such studies have been carried out in the areas of educational finance and enrolment.

In addition, the OECD runs trend data collections every year to re-collect data for past years on a consistent (similar) basis approach, in order to have comparable data over time and ensure that any adjustments to previous data have been taken into account in the most current data collection.

6.4 Suggestions for the estimation of missing data

National data sources are rarely adequate to provide all of the data requested at the international level and missing codes frequently have to be used. This section provides some suggestions on techniques that can be used to derive estimates for some of these missing values. In each case they are, merely suggestions; the data providers are best placed to judge how reasonable the estimation techniques are in their own countries' data.

There are broadly five situations in which missing values might arise:

- **Data not collected for a variable**. In this case, it may be possible to create an estimate based on assumed relationships to other variables. For example, if students' age distribution is not available but the grade distribution is, it may be a reasonable assumption that all students in the same grade are the same age. Alternatively, there may be information about the relationship between age and grade from another source (such as a research study or ad hoc survey) which can help with estimating the missing variable.
- **Data not available for the desired level of aggregation**. A common example here would be where data only provide partial national coverage, e.g. are available for some regions but not all. Here a feasible approach may be to scale up the subnational figures to national level using a scaling factor derived from a different, but related dataset. For example partial student enrolment numbers could be scaled up on the basis of student data from labour force surveys or from the results of an ad hoc survey.
- **Data only available for certain sub-populations**. This case is similar to the previous situation and the same potential solution could be applied. For example, where certain data may be available for public schools and government-dependent private schools but not for independent private schools, they could be scaled up as described above.
- **Data not available for the desired level of disaggregation**. For example, expenditure data may not be available for each level of education separately but can be apportioned to the corresponding levels based on student enrolments in the respective levels. Alternatively, expenditure could be apportioned based on the relative student-teacher ratios between the levels, or staff numbers. Similarly, teacher numbers or teaching hours could be used to distribute teachers' salaries between ISCED levels. A related situation is where most national data can be allocated to the international classification but there are a number of cases that cannot and would otherwise be recorded as "not known". Here, the "not knowns" could be allocated to the target classification on a pro-rated basis.
- **Data not available for the year of the data collection**. In this case it may be possible to estimate the data on the basis of data from previous years. For some finance data, applying inflation rates to a previous year's data may be appropriate as long as that is seen as a reasonable estimate of the expenditure that will actually have occurred. Budgeted rather than actual expenditure figures may also provide a reasonable basis for estimating current year expenditure. For student enrolment data, current year estimates could be derived by applying estimates of transition rates between levels or grades, preferably based on historical trends.

In all cases, when choosing a technique to estimate missing data, thought needs to be given to the use to which the data will be put, particularly in indicator calculations. For example, using student numbers as a basis for estimating missing expenditure data would be inappropriate if the estimated expenditure data were then to be used to calculate expenditure per student.

6.5 Remaining areas for data-quality improvement

Although much progress has been made in improving the comparability of international education statistics and indicators, much has still to be done. Comparability could still be improved in the following major areas:

Coverage of educational programmes

Although non-formal education is a recognised part of the international classification of education, as defined in Chapter 3, international data collections are likely to restrict their coverage of educational statistics to formal programmes for the sake of international comparability and feasibility.

The heterogeneity of non-formal education programmes means that it is difficult to provide general guidelines for their application in statistical instruments, given the purpose of international comparability. Currently, the OECD recommends using the criteria of equivalency of content for the classification of non-formal education programmes, which relate non-formal programmes to formal programmes with similar content within ISCED. However, at this stage, ISCED 2011 does not give specific advice on the development of mappings for non-formal programmes or any related non-formal educational qualifications.

Classification of programmes by level

According to the ISCED manual, the notion of "levels" of education is represented by an ordered set, grouping education programmes in relation to gradations of learning experiences, as well as the knowledge, skills and competencies which each programme is designed to impart. The "level" reflects the degree of complexity and specialisation of the content of an education programme, from foundational to complex. However, curricula are too diverse, multi-faceted and complex to directly assess and compare the content of programmes across education systems in a consistent way. In the absence of direct measures to classify educational content, ISCED employs proxy criteria. These proxies only provide a pragmatic answer and efforts need to continue to arrive at a more comparable allocation of programmes to levels.

Full-time and part-time student status and conversion to full-time equivalents

The reporting of these data to common international data definitions is one of the areas that is most constrained by what is collected nationally. As noted in Chapter 4 (Section 4.1.9), up to the end of secondary level, the method used to distinguish between full-time and part-time students is more likely to depend on student attendance or time in the classroom. At tertiary level, study load is more likely to be measured in terms of instructional hours and credit accumulation, but this may not be consistent across countries. Moreover, some countries distinguish between full- and part-time on the basis of the characteristics of the programme rather than of the time students spend studying. For instance, in the particular case of combined school and work-base programmes, students participating in these dual-system apprenticeship programmes are classified as full-time students even though the school-based component comprises only part of the programme.

In addition, the factors used for converting these student numbers to full-time equivalents will not necessarily be derived on the same basis. Some will be based on classroom attendance, some on study time commitment and some on credit accumulation, and this is likely to lead to some distortion in international comparisons. The indicators affected will be those on ratios of students to staff and expenditures per student.

Successful completion/graduation

The recent revision of the ISCED classification helped to clarify international definitions of graduation. When a qualification obtained does not provide direct access to a higher ISCED level, successful completion

of programmes may be considered as level completion (without access) or no level completion. If such a programme meets the right criteria, completion could be partial (more details in Chapter 4, Section 4.1.4). The inherent difficulty lies in being unable to measure the quality or value of a graduation across (and within) countries. This would require an international standard or benchmark which is not available at present.

Ancillary services expenditure

While it is clear that expenditure on ancillary services within educational institutions should be included in the reported data (see Chapter 4, Section 4.5.3), the extent to which they are varies from country to country. Where countries do report such expenditure, it remains difficult for many of them to report it separately from expenditure on educational core services, particularly at the tertiary level. This could lead to distortions in the expenditure indicators and prevents these indicators – particularly expenditure per student – from being calculated on a more logical basis using core services expenditures only.

Financial aid to students

Generally there is a need to seek fairer and more complete measures of the financial aid provided to students. Two issues in particular have not yet been adequately addressed: first, the reporting of student loans and second the tax benefits and allowances paid to students and their families which are contingent on the beneficiary being a student (see Chapter 4, Section 4.5.4). Student loans are currently measured on a gross basis, without subtracting repayments. While this is acceptable as a measure of the financing of students in the current year it does not adequately measure the generosity of the aid package available to students and nor does it fairly reflect the share of cost between the public and private sectors. Tax benefits to students and their families are excluded from the expenditures on education as there is no internationally agreed methodology for measuring and reporting them, and yet these are legitimate means of providing support to students and their families. Excluding such expenditure therefore undermines comparisons of financial aid to students and of public subsidies to households generally.

Student mobility

Mobility measurement in education has gained importance in the last years, which translated to an effort to better define what exactly is covered (Chapter 3, Section 3.3.5). Henceforth, efforts are needed to capture better data to improve the comparability of the foreign student data and differentiate foreign from international students.

References

OECD (2012), *Quality Framework and Guidelines for OECD Statistical Activities*, Version 2011/1, Statistics Directorate, OECD, Paris, www.oecd.org/statistics/qualityframework.

UIS, OECD and **Eurostat** (2016a), *UOE Data Collection on Formal Education: Volume 1, Manual on Concepts, Definitions and Classifications*, UNESCO Institute for Statistics, OECD and Eurostat, Montreal, Paris, Luxembourg.

UIS, OECD and **Eurostat** (2016b), *UOE Data Collection on Education Systems: Volume 2, Questionnaires and Instructions for their Completion and Submission*, UNESCO Institute for Statistics, OECD and Eurostat, Montreal, Paris, Luxembourg.

United Nations (1994), *Fundamental Principles of Official Statistics*, Statistics Division, United Nations, https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.