Comparability Issues in International Educational Comparisons

Andreas Schleicher

OECD, Paris

Abstract

The diversity of education systems and differences in the structure of the governance of education make international educational comparisons very difficult. There are various parameters that affect the comparability of international educational data and data which are adequate for certain types of comparisons may reveal to be entirely inadequate for other types of comparisons. This article attempts to identify these parameters and illustrates the impact of methods of data collection on the comparability of data. Account is thereby taken of various trade-offs between comparability, coverage, uniformity, complexity, and other attributes of educational comparisons. This article concludes that there are no absolute criteria for comparability but that for an adequate use of data the knowledge of different quality-relevant properties of the data in the form of structured meta-data is an essential requirement.

1 Introduction

Rapidly changing profiles of labour markets and educational demands have generated a growing need for valid and comparable educational data that can assist policy makers to plan and manage the supply of educational services and to monitor educational progress. Data are needed that provide insight into the components of the education systems and their interrelationships. For example, data are needed on the conditions under which the education systems operate in order to account for the "givens" (e.g. demographic and socio-economic characteristics of the population, special student groups, financial resources available, and public support). Similarly, data are needed on the processes and programme features that reflect the decisions, policies, and practices of the education systems (e.g. participation and flows, decision making characteristics, human resources, financial resources applied, schooling processes, staff characteristics, and instructional characteristics). And finally, data are needed

on student outcomes, system outcomes, as well as labour market and social outcomes so that the results achieved by the education system can be evaluated.

For practical and ethical reasons, the necessary data cannot be obtained from large scale experimental studies in which the introduction of "treatment variables" occurs according to a pre-arranged experimental design and all extraneous variables are either controlled or randomised. Hence policymakers often rely on the variation that exists between the education systems of the different countries. This is why international comparisons play such an important role.

However, differences between the education systems and the structure of the governance of education make such international educational comparisons very difficult. Even if data are reasonably accurate and adequate for the needs of national data requesters they may not be comparable at an international level because of, for example, differences in national definitions and classifications. Further complications arise from the fact that education systems cannot be held to a fixed position and it is thus difficult to ensure the validity of comparisons where the framework of the education systems and the policy priorities change over time.

A common approach to ensure international comparability in such situations is to narrow the objects of comparison to a common international denominator and to establish easily quantifiable criteria of comparison (for example, in evaluating education achievement, one may assess only students that share one common characteristic, such as a particular age).

On the other hand, the striving for an exact homogeneity of the populations that are compared and the application of very restrictive methods of comparisons can limit the scope, validity and usefulness of the collected information. It is therefore necessary to take into account various trade-offs between comparability, coverage, uniformity, complexity, and other attributes of educational comparisons. It is also important that the concern for accuracy and comparability in easily quantifiable terms does not obstruct the production of more complex data which are less easily simplified and standardised but which may make important substantive contributions.

The *key comparability requirement* is to ensure that the errors that are introduced by comparability problems are small as compared with the differences that the comparison is intended to explain. For example, if a comparison reveals that two countries differ in the achievement of their students in mathematics by 10 percent points on an international scale it must be ensured that potential sampling, non-sampling, and measurement errors are small compared to this difference.

This article addresses issues of the comparability of educational data in several stages. First it identifies and illustrates a set of core parameters that are relevant to the comparability of educational data, and afterwards it illustrates the impact of methods of data collection on the comparability. It is shown that there are usually no absolute criteria for comparability but that for an adequate use of data the knowledge of different quality-relevant properties of the data is essential. This leads to the need to associate data about "real world objects" that are collected with various types of meta-data which can then guide the evaluation of the adequacy of the data for the specific objectives of the comparison.

2 Parameters Relevant to the Comparability of Data

Certain groups may be sufficiently similar to permit valid comparisons between them even though individual characteristics of the populations making up these groups may very widely. When establishing criteria for comparability it is important to define carefully the object types and populations that are being compared and from which data are collected, the classification criteria or dimensions that form the subgroups of comparison, the variables that are measured, and the methods of comparison that are used. These parameters are described below and illustrated in a number of examples.

2.1 Comparability of Object Types

Each data element is collected with an explicit or implicit reference to an *object type*, denoted in the following by $O(t_o)$ which exists during time t_o. Object types can be students enrolled, entrants to or graduates from educational programmes during the school year t_o, or they can be classes, types of educational institutions, or entire components of education systems. For example, a data collection may focus on the comparison of student achievement, on the comparison of the effectiveness of schools, or on the management of resources in particular schooling systems.

A comparable definition of the object types is usually not trivial. For example, when data are collected on education personnel, careful consideration must be given on whether data are reported as head-counts, as full-time equivalents, or as posts or jobs (where a job is understood here as an implicit or explicit contractual relationship between a specific person and a specific

post). As another example, if data are collected on entrants or new entrants to a specified level of education then clarification is required on the types of programmes that lead to a recognised qualification at this level of education as well as on the different types of student flows that are classified as entrants or new entrants. For example, should returnees and re-entrants to a programme or level of education be included or not? How should enrolees who enrol in a second or further programme at a given level be treated? Similarly, if data are collected on graduates the criteria for "success" need to be well defined and operationalised: In some countries this can be linked to the obtaining of a degree or diploma after a final examination while in other countries, it can be defined by the achievement of programmes without a final examination.

Each educational comparison is usually based on a single object type even though a data collection as a whole may well serve different types of comparisons each involving different object types. For example, in a data collection which collects data from students this information may be aggregated to the class or school level or to even higher levels of aggregation.

The objective of the comparison is to identify differences of the objects with respect to certain well defined variables (denoted by *V* in the following) while the objects that are being compared are believed to be either similar with respect to certain other parameters or in which these other parameters are believed to be controlled or randomised. It is usually this second set of parameters that is of concern when issues of comparability are discussed. In the real world it is usually neither possible nor useful to strive for an exact similarity of homogeneity of the groups that are being compared but the objective must rather be to distinguish those parameters that are similar between the population elements from those that are not. Therefore, comparisons between "apples and oranges" are not necessarily useless as long as we know which are the characteristics that are shared between them and in which they differ. For example, depending on the policy questions to be answered, in one situation it may be appropriate to compare students of a given age whereas in another situation it may be more appropriate to compare students in a given grade, and yet in another situation it may be most appropriate to compare students who have been subject to a common curriculum. Yet a survey that is designed to compare the achievement of students of a certain age between country, may yield incomparable results with respect to the grade, the curriculum, etc.

While a comparable definition of the object types may still be relatively straightforward when the object type refers to individuals, it usually poses substantial problems when the objects are defined at an institutional level.

Consensus is then required on the criteria that will be used to define such units. For example, specification is required as to whether a unit reported as a university is defined geographically, from a management perspective, based on certain characteristics of the performing institutions, from the perspective of the principal activity, etc. Once these criteria are established conceptually, they need to be operationalised and many questions need to be answered: For example, how and by whom is management and control exercised and how is it shared with higher levels of educational authorities? Should it cover the determination of curriculum and course content; the selection and employment of teaching and other staff; the determination of the intake of students; the determination of day-to-day spending? How are "off-campus" centres of the university treated? Or how are centres that are dedicated to education or training but exists within a non-educational institution (e.g. special training units within firms) treated? Should the coverage be restricted to institutions providing directly or indirectly instructional services to students or should institutions that provide administrative and support services be included as well? Should centres of purely academic and applied research and administrative agencies be included that are not engaged in teaching and that do not control any teaching institutions?

We can formalise the measurements or observations as an ordered pair $< O(t_o), V(t_v) >$ where t_o refers to the time during which the object type exists and t_v refers to the time of measurement. When we draw a sample of students and, for example, administer a test to the students in the sample, we will obtain responses or *values* v_j of the variables V from particular students which are referred to as *instances* o_i of the object type O "student". Such data are also referred to as *micro-level data*.

However, such micro-level data are usually of little interest in themselves as they do not allow generalisations to be made. To obtain useful information these micro-level data need to be aggregated along certain classification criteria to so-called *macro level data*, in every-day language also referred to as *education statistics*. When comparing education statistics we deal with the results of estimations of a set of statistical characteristics (where the estimations are made on the basis of a set of micro-level data).

A formal description of macro-level data becomes more complex because there are many more parameters involved than in micro-level data; These include the object types which comprise the target population, the type of classification variables that are used, the time at which measurements are undertaken, the variables which are measured, and the statistical estimator that is used to summarise the data. All of these elements affect the comparability

of the resultant data and need to be taken into account as shown in the following.

2.2 Comparability of Target Populations

When making comparisons of macro-level data, usually only those instances of the object types will be included in the comparison which hold certain specified attributes or fulfil certain specified criteria. For example, if the object types are students, the definition may require only those students to participate in the comparison who share a common age, or the enrolment in a common grade, type of educational programme, or type of education service provider. In the following, the vector of attributes that defines the target population will be denoted v_o. The resultant set of objects is usually referred to as the *target population* - denoted by $O(t_o)$ (with v_o) - and operationalised through the definition of its content and extent, the definition of the units of sampling and analysis, and the time to which the definition and data collection refers.

Comparisons of macro-level data thus not only require the comparability of the object types, but additionally require the comparability of the entire target population. This may be straightforward in some educational surveys where the target population can be specifically and narrowly defined so as to match the purpose of the survey. However, in general, and especially in situations where comparisons are based on existing national data, this can pose substantial comparability problems.

A comparisons of education statistics relies first of all on the assumption of a similar coverage of the statistics in the education systems involved. To assure this, reference is frequently made to a conceptual definition of education. For example, reference is often made to the International Standard Classification of Education where education is defined as "organised and sustained communication designed to bring about learning" (UNESCO 1976).

Though it may be relatively easy to obtain international consensus at this conceptual level, the operational definition of the terms used in such a definition is usually far less clear and often requires a careful enumeration of the boundary cases. When comparing data on participation, for example, basic criteria need to be established for the functional and institutional distinction of the education sector from the wide range of related sectors which may be very closely related to it in terms of functions, institutions, and personnel. The following outlines just a few cases from the whole range of borderline cases for this problem.

For example, what are the boundaries between education and child care, and specifically, where can the starting point of pre-primary education be defined as the point where the teaching function takes precedence over the childminding function? This can have a considerable impact on statistics on expenditures per pupil at the pre-primary level because the extent of non-educational components at the pre-primary level of education varies considerably between countries. Another example is the boundary between education and research and development. In institutions of higher education, research and training are always very closely related. Because the results of research feed into teaching, and because information and experience gained in teaching can often result in an input to research, it is difficult to define where the education and training activities of higher education staff and their students end and research and development activities begin.

Substantial differences also exist between countries in the treatment of informal education, adult education, education outside institutional settings, education provided by the enterprise sector, leisure and culture. In order to compare the levels of participation adequately the coverage of these forms of education needs to be precisely and operationally specified. For example, the requirement that "adult education" should be covered if it is similar to "regular" education will only carry meaning if countries share a common understanding of what "regular" education is, and what the criteria of similarity are. The latter, for example, can refer to very different aspects such as the similarity of the studies, the similarity of the educational programmes, the similarity of the potential qualifications of the programmes, the similarity of the consumers of education, etc.

Finally, countries have various programmes and delivery mechanisms for educational services for mentally, physically, or emotionally disadvantaged students and for other groups with special learning needs. Operational criteria need to be found that align the coverage between countries. For example, practices vary in terms of definitions, programmes offered, the degree to which special education is integrated into the regular education system, the classification of special education students, and the type of support given to these students.

2.3 Comparability of Classification Categories

The objective of a comparison is usually to compare measures between subgroups of a target population which reflect certain structures of the target

population. The criteria that are used to establish these subgroups are referred to as *classification criteria* or *dimensions* and will in the following be denoted as a matrix $C(t_c)$ where t_c is a vector of time parameters that corresponds to the vector of classification variables (for example, if we classify enrolments by age, the time parameter could refer to the anchor age of the student ages). The specific categories of the classification criteria are referred to as *classification categories*. For example, if the variable of interest is the number of students enrolled, students can be cross-classified by the level of education, the type of educational programmes, and the type of service provider in which they are enrolled. Similarly, it is possible to classify educational expenditures by nature of expenditure, type of transaction, and type of educational institution.

The same type of comparability issues that exist for the comparison of target populations now arise for comparisons between these classification categories. For example, if the classification criterion is the type of educational institution, comparability problems arise when the terms used to characterise the classification categories do not share a common meaning across education systems. The distinction between "private" and "public" schools, for example, may be well-defined within a certain education system so that it is possible to classify each school in the target population in that country unambiguously as either "private" or "public". However, these terms could have a different meaning in different countries and thus have no international validity: In one country, the distinction between public and private schools may refer to the management control; in another to the extent in which funds are derived from public or private sources; and in yet another to the degree of institutional dependence on funding from government sources.

Another example for comparability issues in classification criteria is the classification of students as full-time and part-time students and the reduction of such data to full-time equivalents: Some countries, for example, base the part-time/full-time classification on the actual time of student participation whereas other countries use it as an attribute of the educational programmes or the provision of education in general (then consequently classifying all students attending full-time programmes as full-time students even though these students may actually devote a substantial portion of their time to other activities). Major differences also exist between countries in the criteria that constitute full-time participation. Since theoretical and typical daily durations of education programmes differ widely between programmes and countries and since there are no international accepted norms, relative national norms are often applied for establishing full-time participation (based, for example, on the percent of

the school day or week as locally defined) which can affect the international comparability of the statistics considerably.

2.4 Comparability of Variables and Measurements

The last set of parameters which need to be taken into account when evaluating the comparability of educational statistics is the vector of variables *V* which are being measured at time t_v and the statistical estimator *f* that is being used to summarise the values *v* of *V* for the instances *o* of the object type *O*. Variables may be the responses of students to a set of achievement items, head counts, a measure of full-time equivalents, etc. The statistical estimators may range from simple descriptive statistics to complex structural relationships.

The variables and the methods of observation may introduce comparability problems of various types. For example, comparability problems may be introduced by non-observational errors such as errors related to non-coverage, unit non-response, or item non-response, or they may be introduced by observational errors such as errors associated with interviewers, the data collection instruments, the data providers, or the data collection operations.

An example of such problems is the distortion of statistics on education expenditures for education personnel by non-comparable methods for the measurement of retirement expenditures for education personnel: Some countries measure the contributions flowing into pension funds whereas other countries have unfunded pension plans and therefore measure expenditures in terms of pension payments made to former employees who are currently retired. The different measurement methods can implicate differences in the resultant statistics that can be of similar magnitude to the differences in education expenditures and therefore threaten the validity of the comparisons critically (see also Barro 1994). Another example is that to differences in reporting participation data: For example, some countries may report in their enrolment statistics the number of students enrolled on a given date which are then used as a proxy for enrolment over the entire school year; others may report the average number of students enrolled during the (calendar) year; and yet others may report the total number of students enrolled during the (calendar) year (thus potentially double-counting multiple entrants and re-entrants). The reference time in the school year when such statistics are collected also differs usually widely between countries.

Finally, when using macro-level data, careful account must be taken of the aggregation function *f* that was used to derive the macro-level data from the

micro-level observations. The calculation of a student-teacher ratio may, for example, yield entirely different results when they are calculated based on data at the student level than when they are derived from class-level or school-level information.

2.5 Formal Structure of Education Statistics

Taking all the above into account, we can formalise the structure of macro-level education statistics as an ordered pair:

* $< O(t_o)$ (with v_o), classified by $C(t_o)$, $V(t_v)$, $f >$

where:

O is the object type which is being measured (e.g. "students", "graduates", or "entrants") which exists at the time or during the period t_o (which could, for example, be a school year, a calendar year, or a financial year).

C is the vector of classification variables.

V is the vector of variables which are being measured at time t_v.

f is the statistical estimator which summarises the values of the variables *V* for the instances of the object type *O*.

As illustrated above, all the elements in this ordered pair need to be considered when evaluating the comparability of education statistics.

3 Methods of Data Collection and Their Relationship to the Comparability of the Data

Comparability issues are further related in various ways to the methods used for collecting data (see also Schleicher 1994). Furthermore, the type of data source needs to be considered when the adequacy of data for a particular comparison is evaluated since different types of comparability issues arise depending on whether the data come from labour force, household, school surveys, population census data, fiscal and national accounts data, administrative records, legal or other documents, or from expert judgements.

To illustrate the potential impact of the methods of data collection on the comparability of the data, this article contrasts the collection of data with reference to an entire target population in the form of a census with the selection of a probability sample from the target population in which every

element of the target population has a known and non-zero probability of selection.

Whenever inferences about individual elements of a population are required it is, of course, necessary to obtain data from each member of the target population and therefore to undertake a census. However, when only inferences on entire subclasses of the population (or the population as a whole) are required then sufficiently accurate results can often be obtained from a relatively small body of data based on only a small fraction of the population. Consequences are reductions in the costs of data gathering, coding, data management, and data analysis. In addition, reduced costs are associated with the training of fewer personnel to conduct the fieldwork.

It must also be taken into account that census data are rarely collected for international comparisons but are usually collected for specific national administrative purposes which may significantly differ between countries. Thus, the underlying definitions that are also used in each country for the data collection may differ widely and thus introduce comparability problems at the international level. In an international survey it is much easier to apply a set of definitions that are jointly agreed upon by the different countries and similarly, it is much easier to adapt the data definitions and requirements precisely to the specific purposes of the comparison. This is of particular importance in international comparisons of educational outcomes.

The reduction in the amount of data obtained from a sample also permits the diversion resources into improving the comparability of the data. More time can be spent on the training of fewer and often higher qualified personnel, and a more intense supervision of the fieldwork is possible. More care can be spent on obtaining the required responses from the sampled observations and to tracing possible non-respondents, thus reducing potential sources for non-response bias. More resources can also be spent on the gathering of more complex information from the respondents. This is, for example, the case when data have to be obtained through interviews or through open-ended or free-response tests which need to be rated and scored. In other cases complex methods of analyses are undertaken that may require expensive technical equipment. Both highly qualified personnel and technical equipment are usually limited so that surveys with more complex methods can often only be undertaken on the basis of a sample.

Finally, because of the smaller number of observations the time required especially for the field administration and data processing can often be substantially reduced in a sample. This is often an important aspect of data

collection thus ensuring the actuality of the results demanded by policy makers and funding agencies.

Certainly a census has the advantage that estimates are obtained without sampling errors. On the other hand, the sampling errors are often small when compared with other errors that occur in a census, such as non-observational and observational errors, methodological errors; finally, sampling errors may be small when compared with changes in the survey variables over time.

4 Meta-Data as a Means to Document Comparability

The preceding sections of this article have briefly described the various parameters that affect the comparability of international educational data. They have further shown that data that are adequate for certain types of comparisons may be shown to be entirely inadequate for other types of comparisons. It is therefore often not helpful to speak of "comparable data" or synonymously "good data quality" as a general attribute of data since this assumes that there is a single and generally agreed upon parameter that determines what "good quality" means. Instead it is important to identify and document the different quality relevant properties of the data that can affect its comparability.

This requires the association of the data about the "real world objects" (such as students or educational institutions) with information on the content and context of the data itself as well as on the processes associated with the production and analytical use of the data. Such information would comprise data sources (instances of this class could be, for example, the source publication or database, the source agency, the reference period and date of publication of the source), content-related information such as international definitions, instructions and recommendations, and information about how the national data providers have interpreted definitions and allocated national categories to international classifications as well as information on the processes employed in measurement. Finally, it comprises information on the methodology used such as how missing data have been treated. Such data are often referred to as *meta-data*.

There are different types of users of educational data which require very different types of meta-data. Some may just need global declarative information in order to assist them in the search for data that can address their policy questions. Data producers may require detailed information on the statistical units, the coverage, and the classification categories of the data. Furthermore those who make use of the statistics for indicator calculations and reporting may

need information on how to interpret and analyse the data and on where countries deviate from the international norms and they need to know which annotations to prepare for which tables in a publication.

Expression (*) provides an indication of the attributes of education statistics with which meta-data need to be associated:

First, meta-data need to be associated with the time parameters t_o and t_v. t_o is often defined as a school year for which data are collected (for example the school year in which a student achievement test is administered), but can also be a financial year, or a calendar year. t_v represents the date or period in which students were assessed or counted (for example, a test may be administered in the second week of the eight's month of the school year, in which case t_v will depend on the beginning of the school year in each country). Often the time parameters t_o and t_v differ between different data series and there is often considerable between-country variation in the time-lag between t_o and t_v. Analysts need to take this into account when interpreting the data. For example, if an achievement test is administered at a certain point in the calendar year, then this means that it is administered in different countries at different points in the school year which can affect the achievement considerably. Similarly, if data associated with a school year are related to data associated with a financial year, the time lag between both must be taken into account. Furthermore, meta-data also need to be associated with the type of statistical unit *O* (and with particular instances of the object type o_i) as well as with the coverage of the statistics as expressed through v_o. Also the classification criteria C and their classification categories (e.g. international changes in definitions or recommendations or national deviations in the implementation from the international definitions) need to be documented. Finally, meta-data need to be associated with the variables *V* that are measured.

The organisation of meta-data is often far more difficult than the organisation of the data themselves and only an efficient and well structured organisation of meta-data will make this information accessible and therefore useful to the different types of users. It is useful to distinguish between three different layers of association for meta-data: One layer that documents the comparability with respect to the above attributes of education statistics, a second layer that provides documentation at the series level, and a third layer that relates to individual occurrences of data elements (see also Sundgren 1994). Some meta-data are related to comparability in time whereas other meta-data are related to comparability in "space". Space in this context is a generic concept, covering not only the geographical dimension but also many other classifications where

it is useful to establish some kind of proximity between different instances of one of the above attributes or between countries.

Proper knowledge and treatment of meta-data can make the difference between valid and misleading data calculation and interpretation.

5 Conclusion

This article has discussed various aspects of the comparability of international educational data, taking into account the various trade-offs between comparability, coverage, uniformity, and complexity. It has shown that there is usually not a single and generally agreed upon criterion that would determine whether data are "good" and "comparable" but that there is a fairly large number of parameters that affect the comparability of educational data, formalised in expression (*).

In an international context where comparative data are often derived from existing national data sources it is usually not possible to ensure comparability through narrowing the objects of comparison to a common international denominator and through the establishment of easily quantifiable criteria of comparison. In such situations only the availability of well structured meta-data on the different quality-relevant properties of the data can help to assure the comparability of the data and the validity of the interpretation.

Bibliography

Barro, S. (1994). *Expenditure Comparability: The Problem of Measuring Retirement Expenditures for Education Personnel*. Technical note presented at the meeting of the OECD/INES Technical Group, Paris December.

Schleicher, A. (1994). International Standards for Educational Comparisons. In A.C. Tuijnman & T.N. Postlethwaite (Eds.), *Monitoring the Standards of Education* (pp. 229-247). Oxford: Pergamon.

Sundgren, B. (1994). *Statistical Meta-Data and Metainformation Systems*. Working paper No. 15 on the Conference of European Statisticians, Statistical Commission and Economic Commission for Europe, Geneva 22-25 November.

UNESCO. (1976). *International Standard Classification of Education*. Paris: UNESCO.