HEC MONTRÉAL

Explorative analysis of patterns and causes of missing data in the UIS educational data

by

Miguel Alberto Ibáñez Salinas

Sciences de la gestión
(Intelligence d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M.Sc.)

Janvier, 2012

HEC MONTRÉAL
affiliée à l'Université de Montréal

DÉCLARATION DE L'ÉTUDIANTE, DE L'ÉTUDIANT
ÉTHIQUE EN RECHERCHE AUPRÈS DES ÊTRES HUMAINS

Recherche sans collecte directe d'information

Cette recherche n'impliquait pas une collecte directe d'information auprès de personnes (exemples : entrevues, questionnaires, appels téléphoniques, groupes de discussion, tests, observations participantes, communications écrites ou électroniques, etc.)

Cette recherche n'impliquait pas une consultation de documents, de dossiers ou de banques de données existants qui ne font pas partie du domaine public et qui contiennent de l'information sur des personnes.

Titre de la recherche : Explorative analysis of patterns and causes of missing data in the UIS educational data

Nom de l'étudiant : Miguel Ibanez Salinas

Signature :

Date : 2012-01-04

Imprimer

Summary

The UNESCO-UIS education database can be considered one of the most comprehensive education database in the world, containing a wide–range of comparable statistics for more than 200 countries and territories, upon which governments, organizations focusing on international development and researchers around the world rely for monitoring education. As a consequence, maintaining and/or increasing its quality are activities of critical importance for the Institute of Statistics (UIS).

A vital quality dimension of this database is its capacity to present complete information regarding education. To address this issue, this exploratory study examines the UNESCO-UIS database from the point of view of the missing values and completeness for data related to academic years 1999 to 2008. It seems that no study has previously been conducted on this matter.

The analytical tools used in this study are multiple (e.g. descriptive analysis, control charts, linear regression, binary factors analysis, cluster analysis, multinomial random effects logistic regression, etc.), a fact that reflect the complex nature of the patterns of completeness/missingness of the education database.

This study begins by considering the negative effects of missing values in the production of statistics, showing that they can increase processing costs and decrease the reliability of inference based on UIS datasets. Preliminary descriptive analyses showed that certain groups of variables may be losing data progressively (decreasing responses across time). Furthermore, more advanced analytical tools allowed us to propose an underlying structure of five dimensions or factors (linked to specific parts of the UIS data collection questionnaires) that describes the manner in which countries' education data are produced. In turn, this proposed structure proved valuable in the classification of countries in five clusters, where each cluster can be linked to the capacity of countries to produce/report data. Finally, the behaviour in time of these factors was analyzed; it was noted that the reports of detailed statistics in primary/secondary and tertiary education statistics are decreasing in time. Other interesting results are related to the link between improvements in governance indicators and the increase in the production of education statistics.

These conclusions have direct implications in the activities of data collection and statistical capacity building carried by the UIS, for example, in the construction of diagnostics tools, country level reports, etc.

Sommaire

La base de données d'éducation de l'UNESCO-ISU peut être considérée comme l'une des plus complètes dans le monde, contenant un large éventail de statistiques comparables pour plus de 200 pays et territoires, et sur lesquelles les gouvernements, les organisations axées sur le développement international et les chercheurs du monde entier comptent pour surveiller l'évolution de l'éducation. En conséquence, le maintien et/ou l'amélioration de sa qualité est une des activités d'importance critique pour l'Institut de statistique de l'UNESCO (ISU).

Une dimension essentielle de la qualité de cette base de données est la capacité de présenter une information complète à propos de l'éducation. Il semble, néanmoins, qu'aucune étude antérieure n'ait été menée concernant cette affaire. Pour répondre à cette question, cette étude exploratoire examine la base de données UNESCO-ISU à partir du point de vue des valeurs manquantes et la complétude des données relatives aux années scolaires 1999 à 2008.

Les outils analytiques utilisés dans cette étude sont multiples (par exemple, l'analyse descriptive, les cartes de contrôle, la régression linéaire, l'analyse factorielle de données binaires, l'analyse de regroupements, la régression logistique multinomiale avec effets aléatoires, etc.), un fait qui reflète la nature complexe du schéma des données manquantes de la base de l'éducation.

Cette étude commence par examiner les effets négatifs de valeurs manquantes dans la production de statistiques, montrant qu'elles peuvent accroître les coûts de traitement et diminuer la fiabilité des inférences fondées sur des données de l'ISU.

Une analyse descriptive préliminaire a montré que certains groupes de variables peuvent perdre progressivement des données (diminution des réponses à travers le temps). En outre, des outils avancés d'analyse nous ont permis de proposer une structure sous-jacente à cinq dimensions ou facteurs (liés à des parties spécifiques des questionnaires que l'ISU utilise pour collecter des données) qui décrit la manière dont les données sur l'éducation des pays sont produites. À son tour, cette structure proposée s'est avérée précieuse dans la classification des pays en cinq groupes, où chaque groupe peut être lié à la capacité des pays à produire des données. Enfin, le comportement de ces facteurs à travers le temps a été analysé. Il a été noté que le taux de réponse pour les statistiques détaillées en matière de statistiques enseignement primaire/secondaire et tertiaire diminue dans le temps. D'autres résultats intéressants sont liés à la relation entre l'amélioration des indicateurs de gouvernance et l'augmentation de la production de statistiques de l'éducation.

Ces conclusions ont des implications directes dans les activités de collecte de données et le renforcement des capacités statistiques réalisées par l'ISU, par exemple, dans la construction d'outils de diagnostic, de rapports au niveau des pays, etc.

Table of Contents

List of Tables

List of Figures

List of Annexes

Acknowledgements

I would like to express my sincerest gratitude to my supervisor, Dr. Jean-Francois Plante, for his continuous support and guidance through the course of this work. Thank you very much.

Quisiera también agradecer a mi madre, Consuelo Isabel Salinas Castañeda, por su continuo amor y soporte a través de mi vida. ¡Muchas gracias!

CHAPTER 1. Introduction

The production of internationally comparable education statistics is a challenging endeavour both from a technical and a political point of view. Nevertheless, the benefits that evidence-based policies and informed citizen have in a society could be considerable. This chapter briefly introduces the reader to the current state of the international data collection on education, focusing mainly on the UIS data collection - considered one of the most comprehensive education data collections in the world - and to the possible effects that missing values have on the UIS education database. These aspects will expose the necessity for an exploratory study of missing value in the education database.

1.1 Missing values

A missing value is a data point or observation that cannot be used in the normal analysis or monitoring activities. Missing values in the UIS education database can prevent analysts from publishing important indicators, such as regional averages of children out-of-school in a given year, the number of teachers needed to fulfil certain educational objective, etc. Furthermore, every missing value in the education database increases the number of corrective actions, such as estimation, validations of secondary data with country authorities, etc., which in turn increases the costs related to data collection. In addition, high levels of missingness in the education database could greatly affect the reliability of estimations, impacting negatively on the validity of any analysis that relies on UIS education data. As a consequence, the examination of possible patterns of missing values must be an essential part of the data collection activities, both to ensure the efficiency in the production of statistics and to maintain the confidence of the users on the statistical outputs derived from the UIS database. Understanding the patterns of missingness can help in diagnosing of countries with reporting problems, and in resolving chronic problems in data collection. At this moment, there are not publicly available studies on possible patterns or trends of missing values in the international education database. The effect of missing values will be discussed in more depth in the Chapter 2.

1.2 The Institute of Statistics of UNESCO (UIS)

The UNESCO Institute of Statistics (UIS) is the statistical division of UNESCO (United Nations Educational, Scientific and Cultural Organisation). Established in 1999 and fully operational by 2001 in its current location (Montréal), the creation of UIS responded to the need of UNESCO for reinforcing and improving its statistical services in order to meet the increasing demands, by

its Member States (over 200 countries and territories) and the international community, of reliable, high-quality and policy-relevant data related to the relevant fields of this UN organization (education, science and technology, culture and communication) (UNESCO-UIS, 2000).

At the moment of creation, the UNESCO General Conference (resolution 43 by the 30th session, November 1999) gave to the Governing Board of the UIS the mandate of setting up a programme focused on the following priorities and considerations regarding the statistics of its relevant fields (UNESCO-UIS, 2000):

Table 1. Priorities and key considerations in the work of the UIS

Priorities	**Key considerations**
1- The description of the types of statistical data and the group of indicators required at the international level.	- Leverage obtained by the UIS' consultative mechanisms that involve high profile parties.
2- The collection and dissemination of information on education, science, culture and communication.	- Monitoring the increasing demands of information from Member States and the international community, and assisting the use of the data in policy research.
3- The improvement of the statistical capacity in Member States.	- Support of commitments, providing training and advisory services and disseminating technical information.

For 2008-2013, the UIS summarize its mission and priorities in four "main action areas":

"i) the collection and maintenance of international statistics which reflect changing policy and are reliable, internationally comparable and robust, as well as feasible to collect; ii) the production and implementation of new statistical standards, classifications, methodologies, indicators and related documentation; iii) the development of the statistical and analytical capacities of Member States; and iv) the provision of analytical services within the context of the Institute's mission." (UNESCO-UIS, 2007 : 7).

As we can observe, 2008-2013's main action areas are not significantly different from the priorities originally proposed in 1999, except that they seem to be more detailed (for example, the second main action area denotes an elaboration of what was originally portrayed as the

description of statistical data and the grouping of indicators). We can also note that most of the main action areas relate entirely to the methodological framework necessary for the collection of internationally comparable statistics (comprising the development of concepts, indicators, etc.) and the dissemination of a reliable education database.

1.3 International Education Statistics

1.3.1 Comparable Education Statistics

As an UN organization, UIS-UNESCO' efforts towards internationally comparable statistics in education can be placed within the context of the measurement of global development/progress (many times towards internationally agreed goals), which seeks to monitor and compare performances of countries or regions around the world. [Note: for more information on the pursuit of progress by the United Nations, please see de Vries, (2001)].

1.3.2 An international movement for education statistics

As a reflection of the increasing importance that the subject of education is acquiring among different local and global actors, we can enumerate some factors that have significantly contributed to the growing demand for internationally comparable statistics in education and in many ways have shaped their development, dissemination and transformation into policy.

- **Human capital and Economic development:** Investment in human capital has been recognized by policy makers as a vehicle for economic development in the competitive global market and as a mean to increase the quality of life of the population (Postlethwaite, 2004). On one side, increasing interactions and global interdependencies at economic levels have intensified the demand for new and more accurate education statistics which should include nowadays information from other nations (National Research Council, 1995). On the other side, education statistics or indicators about local and international education systems (such as numeracy and literacy levels, enrolment rates, higher and technical/professional education availability, quality of education, etc.) collected in a systematic manner are useful for analysts and policy makers who want to assess the requirements, or comply with the demand, of this competitive "information economy", or simply improve policy-making. In addition, the availability of education data, at national and international levels, demonstrates the commitment of the government and other national actors to human capital growth (or other types of social

capital) and allows comparison of national performance with other trade or investment competitors (Walberg and Zhang, 1998; Kenneth and Jürgens-Genevois, 2006; National Research Council, 1995). Certainly, the increasing demand for reliable data on education systems does not only concern governments (which also invest directly in education), but also private investors and international donors interested in development (National Research Council, 1995).

- **Education for All (EFA):** EFA is a global initiative which unites governments, many UN agencies, several types of governmental and nongovernmental organisations and other international organisations towards the goals of achieving universal basic education, improving educational standards, and eradicating of illiteracy worldwide (McEwen, 1990). Important aspects of the EFA initiative are the significant international aid it helps to direct and the considerable influence on education policy of developing countries (Skilbeck, 2006). Given this, the production of internationally comparable statistics on education is at the core of monitoring the diverse indicators that translate the EFA goals. However, keeping track of the goals is not considered an easy task. Indeed, evaluations of state of education data during and after the first EFA meeting in 1999 made evident the need for further improvements in the quality of statistics - including strengthening the efforts on development (theoretical frameworks and standards), collection (the use of different sources), analysis (results, relationships among indicators, processes, etc.) as well as national statistical capacity building (Skilbeck, 2006). Note: UNESCO is the leading agency of this coalition, and as such, it sustains an international expert team with the task of monitoring and reporting the progress toward well defined operational objectives, with responsibilities stretching over all countries (McEwen, 1990; UNESCO-UIS, 2008).

  Furthermore, there are other international initiatives (e.g. Millennium Development Goals) and international organizations (e.g. the World Bank) linked to social and economic development that face the same problems and share the same concerns regarding international statistics as EFA and UNESCO (for more details, see Heyneman, 2003). As the National Research Council puts it: "The efforts of humanitarian and social justice agencies are often spurred by knowledge of oppressive or dysfunctional conditions. The policies of individual governments, private economic investors, and international development donors are now crucially linked to the availability of information and information systems that cannot only appraise a nation's absolute and comparative status on a particular infrastructure or social capital dimension

but also provide information regarding its progress and performance in building human and other forms of social capital." (National Research Council, 1995 : 35).

- **Construction of a knowledge base on education:** At the foundation of all demands for high quality statistics on education is the concern about increasing the understanding and knowledge of the education systems, of the actions and elements necessary to achieve national and internationally agreed education-related goals, and of the relationship between education and economic and social development (Skilbeck, 2006). Furthermore, there is a special worldwide interest in assuring that national policies are based on evidence or in measures of performances where links among causes and effects are clearly stated and understood (Lewin, 2011). In this regard, the adequate understanding of education systems not only concerns the academic community, but also governmental bodies in charge of defining education policy and the public in general (Kenneth and Jürgens-Genevois, 2006). At the same time, due to their complexity, implications and considerable resources involved, the study or development of indicators can only thrive when the main policy interests and concerns of policy-makers, national administration and education statisticians are aligned to the work of the research community (UNESCO-UIS, 2008).

In conclusion, nowadays, reliable international education statistics are in great demand, and the expectations of the possible benefits for policy-makers, researchers and the general public are higher than ever before.

1.4 Why comparative statistics on education?

One principal characteristic of international education statistics is that, because of standardized definitions and procedures for data collection, processing and analysis, they allow cross national comparative assessments (which are becoming paramount for countries' education policies). As mentioned by OECD (2000 : 5): "A quantitative description of the functioning of education systems can allow countries to see themselves in the light of other countries' performance. Through international comparisons, countries may come to recognise strengths and weaknesses in their own systems and to assess to what extent variations in educational experiences are unique or mirror differences observed elsewhere".

Quantitative comparisons - which, as it will be described in the next section, are based on standardized and agreed concepts and frameworks - of the world's education systems give countries the means to gain knowledge, from one another, on how to spread the benefits of education across their societies, how to strengthen its capabilities to produce a competitive

labour force and how to support lifelong learning through the efficient management of educational resources (OECD, 2004). It has also been argued that internationally comparable statistics facilitate national and international debates about education reform (UNESCO-UIS, 2008). Cross-national research then becomes a laboratory to study variations on quality of education, policies related to education and human capital, best practices, country traditions and other aspects that have a consistent impact on learning and education, while helping us understanding what actions are possible beyond national traditions or laws (Kenneth and Jürgen-Genevois, 2006).

Another important aspect of education development that could greatly benefit from comparative studies is statistical capacity building - the production of national and international education statistics.

CHAPTER 2. Literature Review

How could we address the following questions?

- How many children are not attending school worldwide?
- What is the number or proportion of children that have completed primary school during the last 10 years? Is this quantity increasing? Which countries are at risk of not achieving their goals?
- How many students access primary, secondary or tertiary education across the world?
- How much money is spent in primary, secondary or tertiary education? How different is the level of spending in different regions? (UNESCO-UIS, 2008)

International education data (for example, the UIS education database) allow us to address these and other important questions about the global state of education.

This type of questions, which are normally related to policy issues or concerns from researchers or the public, and other important features of education systems can only be answered through high quality statistics or indicators. At the same time, the negative effect of missing values can be felt across many activities related to the production of statistics, affecting both its costs and the expected quality of its output.

2.1 UIS Data Collection

2.1.1 Definition of an education indicator

There seems to be many definitions for "indicator". The UN-Economic and Social Council remarks that some aspects are common to most definitions and suggests that an indicator is: "a statistic, a fact (quantitative) or encompassing forms of evidence, perception (qualitative); defined for some purpose, such as to assess, valuate, measure, convey a message; reflect some underlying goal, values, conditions, message and so on." (UN - Economic and Social Council, 1999 : 26). De Vries (2001 : 319) notes that the need for an established goal is not compulsory and highlights the fact that an indicator could be "a single number, ratio or another observed fact that serves to assess a situation or a development". Along these lines, Rowe and Lievesley (2002) describes the construction and use of "educational performance indicators" for accountability, monitoring progress, comparison to other systems and political reform, while defining "performance indicators" as "data indices of information by which the functional quality of institutions or systems maybe be measured and evaluated" (Rowe and Lievesley, 2002 : 1).

Indeed, the interest on education indicators goes beyond their arithmetical or computational aspects; education indicators are powerful information tools, whose importance extends well beyond statistical offices. To this effect, Bottani and Tuijnman (1994 : 26) state that: "an indicator is not simply a numerical expression or a composite statistic. It is intended to tell something about the performance or behaviour of an education system, and can be used to inform the stakeholders – decision-makers, teachers, students, parent and the general public. Most importantly, indicators also provide a basis for creating new visions and expectations."

Therefore, based on previous discussions, we could conclude that education indicators are measurement constructs intended to periodically provide stakeholders with quantitative information about the state or different facets of an education system (Bottani and Tuijnman, 1994) and in the case of international education indicators, to provide, in addition, quantitative information that allows monitoring and comparing education systems across the world.

2.1.2 Development of indicators

The development of international indicators is a complex task that involves coordinating the needs of many important national and international education stakeholders and, as mentioned before, it entails more than the technical aspects related to data collection. Indeed, the international education indicator development has been mostly defined as a political exercise, involving in early stages the designation of the political objectives that must guide the selection of indicators and the elaboration of data collection instruments (Cussó and D'Amico, 2005). As stated by Bottani and Tuijnman (1994 : 26): "the development of a set of international education indicators is not merely a technical exercise planned and controlled by statisticians, but first and foremost it is a political one.".

Blank (1993) outlines the steps involved in the development of an indicator system. Although he suggests these steps for the construction of an indicator system in the United States - taking into account federal and state levels of reporting, the flow of needed actions describes a process that is suitable for international education statistics development as well.

As per Blank (1993 : 37), the development of an educational indicator system comprises nine steps:

<u>Selecting indicators</u>

1) Develop a conceptual framework based on research results and the interests of policymakers and educators.
2) Obtain commitment and cooperation of leaders.

3) Involve policymakers, educators, researchers, and data managers in selecting priority indicators.
4) Select a limited number of indicators and minimize complexity in reporting.

Organizing a Cooperative Data System

5) Decide methods of data collection.
6) Work with data users and providers to establish standards for producing comparable data.

Reporting Comparative Data on Indicators

7) Design data forms and cross-walk procedures.
8) Collect and edit data.
9) Report indicators.

The first step, development of conceptual framework, duly reflects the necessity of fulfilling the information requirements from policy makers and other stakeholders. Later on, we will see that this is a critical requirement for having high quality indicators. At the same time, the UIS' publication "Global Education Digest 2008" (UNESCO-UIS, 2008) - an edition dedicated to the UIS data collection - succinctly explains some activities that can be mapped to the suggested nine steps, such as: the conceptual framework for educational indicators, standards for comparable data (classification of educational programmes), data collection procedures, validation of data and dissemination.

In addition, there are some issues to take into account when selecting and developing indicators. For example, OECD (2004) describes three considerations guiding OECD indicator related activities [similar guidelines were noted in UNESCO/OECD World Education Indicators Programme (2000)]:

- Indicators must focus on educational issues where an international comparative perspective adds value over national analysis.
- The indicator development programme must appropriately balance the progress of educational issues for which the data collection is feasible and the stakeholders agree about its utility, compared to areas that need more investment in conceptual and empirical work to increase political and public awareness as well as technical capacity. The feasibility of data collection is primarily related to the ability of countries to produce certain expected or traditional statistics. From the point of view of UIS education data collection, when a country cannot report data through the UIS

standardized education questionnaire, this becomes a problem of missing values, and generally implies the necessity for improvements to the national statistical capacity of a given country.

- Cross-national validity and reliability must be assessed continually.

Similarly, Blank (1993 : 65) proposes three criteria to evaluate and prioritize the development of indicators in the case of national educational indicator systems: "(a) importance/usefulness of the indicator, (b) technical quality of the data (available or expected), and (c) feasibility of obtaining state-by-state data".

These criteria stress the interplay between assuring that indicators are relevant at the international level and assuring that the same indicators are reliable and that the respective data collections are feasible and sustainable across the time (relevance/feasibility and cost/benefit: a relevant indicator may not be consistently produced or it would need great investment to produce, while some easy-to-get indicator may not be relevant).

2.1.3 Statistical Programme at UIS

For Cussó and D'Amico (2005 : 23), UIS statistical programme – the approach taken by UIS to develop and produce international education statistics – is composed of no less than seven parts (freely translated from the original French text) :

- "The definition of political objectives (general and specific);

- The conception of standardized statistical questionnaires, which become the instruments of measure (three questionnaires concerning pre-primary, primary and secondary education, tertiary education and the public and private financing of education);

- The definition of indicators including the creation and provision of a classification system for educational programmes and of methodologies for calculating indicators;

- The provision of training (workshops, seminars) and training material (manuals) with the aim of familiarizing national officials with statistical questionnaires and with the international classification of educational programmes;

- Processing of data collected from the statistical questionnaires, including their verification and analysis using tools of storage, calculation and correction of data - tools related in part to the database.

- Feedback to the national officials if inconsistencies are detected in the submitted data.

- Dissemination of statistics (education database)."

Cussó and D'Amico's description of the UIS-UNESCO statistical programme has certain degree of similarity to the steps proposed by Blank (1993) for building an educational indicator system and to the description of UIS activities made by UNESCO-UIS (2008).

Certain characteristics of the UIS statistical programme merit additional remarks:

- The data sources associated to the production of international comparable educational indicators are multiple. The added value of UIS' statistical programme is the recollection of standardized official statistics submitted directly by Member States. Other important data sources are: World Bank data on economic indicators and the United Nations Population Division (UNPD) data on population; and in a lesser measure: United Nations Statistics Division (UNSD) (for economic data not found in World Bank datasets) (UNESCO-UIS, 2008). These institutions also have the objective of producing international comparable data. The quality of these input data is essential for UIS in order to guarantee reliability and international comparability of its own indicators. As it would be expected, any problem with the quality of the inputs, including missingness (missing values in a source dataset), will affect the production of the UIS educational database.
- The UIS statistical programme is a continuous process, and each element can influence other elements that are not necessarily contiguous in the flow line. The analysis and verification of data could bring light over some issues; in turn, this may start a revision of the definition or the objectives of an indicator. For example, this would be the case when only few countries report a specific data point needed to calculate an indicator. Further internal analysis will have to consider: the original definition of the requested statistic (variable or data point), the added value of this statistic for national authorities, and the state of national statistical capacities regarding the production of the missing statistic.
- The verification and analysis of data (e.g. cross-check of data tables, inconsistency detection and comparison of data trends) reported by national official is an iterative (feedback) process, involving the fluid communication of national official and personal of the UIS repeated times until agreement or expected standards are reached. The quality of international education statistics depends greatly upon the proper data validation made by the UIS. Also, at this point, the concern for missing values in data

tables is evident, initiating the costly correction activities such as estimations, research of secondary sources, etc. (UNESCO-UIS, 2008).

There are multiple aspects related to nature of education systems that are very complex and that certainly influence the choices made in the construction of national education indicator systems. These features are not within the scope of the present research; nevertheless, as a reference, we could cite two of them:

- Education systems are hierarchical organizations where decision-making and accountability are distributed across many actors in the system. Therefore, indicators may be developed to measure characteristic of the nature of the system structure in addition to measuring student's learning processes (Bottani and Tuijnman, 1994 : 24).
- Accountability in education can influence the development of indicators by proposing some priorities or strategies oriented towards: testing and reporting performance of education actors, monitoring compliance with rules or standards, incentive systems, control over schools and management of authorities and teachers (Bottani and Tuijnman, 1994 : 24).

2.1.4 Transition from national statistics to UIS education indicators

As mentioned before, the UIS is in charge of the production of comparable education indicators for over 200 countries or territories (193 Member States and 7 Associate Members of UNESCO) (UNESCO-UIS, 2006). For a considerable part of its inputs (data), the UIS depends on the statistical authorities (ministry of education, national statistical office, etc.) officially designated by Member States to function as respondents of the annual UNESCO education survey, which is the standardized measurement instrument used by UIS to collect national education data (UNESCO-UIS, 2008). Other main sources of data include economic indicators produced by the World Bank and population data produced by the UNPD. The survey's construction and rest of the data collection processes are guided by the UIS statistical framework. This framework determines, based on stakeholders' information needs and technical considerations, the definitions and parameters of the data that need to be collected from Member States. In turn, the statistical framework also guides the conversion (requiring intensive database calculations) of these (national) statistics into international comparable indicators.

Between the collection of inputs (national statistics, other data) and the dissemination of outputs (international comparable indicators), there is an intensive work on the inputs (data processing, data validation), and the verification of the outputs (data estimations in case of missing values,

calculations and validation of indicators). Through these processes, the UIS must guarantee that the data collected and the indicators that depend on these data display the required standards of quality.

The present research focuses on the education indicator dataset, which depends on data that are collected by the annual UNESCO education survey. This survey includes 3 questionnaires:

- Questionnaire (A) on statistics of education - pre-primary, secondary and post-secondary non-tertiary education (with coding UIS/E/reference A, where reference represents the year of collection).

- Questionnaire (B) on statistics of educational finance and expenditure (UIS/E/reference B).

- Questionnaire (C) on statistics of tertiary education (UIS/E/reference C).

A brief list of the educational programmes that the UNESCO education survey includes is given by UNESCO-UIS (2011b : 9):

"The programmes on which data should be reported in these questionnaires therefore include:

a. **regular** education in pre-primary, primary, basic and secondary schools, and in colleges, universities and in other higher education **institutions**;

b. education in **public** (or state) and in **private** schools, colleges or universities;

c. **special needs education** (both in regular schools and in special schools);

d. **distance education** (especially at the tertiary or higher education level);

e. both **full-time** and **part-time** education;

f. the education of international students as well as of nationals or citizens of your country."

National respondents, such as education planning departments, usually publish reports and/or complete UIS questionnaires using three types of statistics: based on sample surveys (household or school surveys), based on censuses (population census), and based on administrative registers (UNESCO-UIS, 2008; Wallgren and Wallgren, 2007). Nonetheless, most of the data collected by the UNESCO education survey are based on administrative records or registers.

Wallgren and Wallgren (2007) mention that government administrative registers are used to store records of all objects within the activity or function that needs to be administrated (e.g. schools, hospitals, etc.). Because these records are created for purposes of administration and identification, it usually corresponds to national authorities to transform these administrative records into statistical records. This transformation renders the statistical records suitable for

statistical purposes. Statistics Canada (2009) cites some examples of the possible uses: for data collection (exercise related to the report of education statistics to UNESCO), for survey frames, for support to edition, imputation and calibration of estimations, and for survey evaluation (comparison of estimates from administrative-based data and survey data).

At the same time, statistical registers exist within a system of statistical registers, usually maintained by national statistical offices or other organizations, such as, ministries of education, finances, etc. (Wallgren and Wallgren, 2007). In the case of administrative data reported in the UNESCO education survey, data are usually collected at the school level, and then aggregated (or transformed) at local, regional and/or national government levels usually by planning departments in ministries of education and or national statistical offices (UNESCO-UIS, 2008).

Wallgren and Wallgren (2007 : 5) describe the possible activities involved in the transformation of administrative records to statistical records (e.g. data editing, variable coding, management of object and variables, handling of missing objects and missing values, etc.). The quality of the statistical registers is also a subject of concern for national authorities as the activities concerning data quality management are interrelated to the transformation of administrative records into statistical datasets. Some important activities related to quality are: contact with data suppliers for further details or corrections, checking of the received data and missing value analysis. As mentioned before, any problem with the quality of educational data at this stage will certainly reflect on the quality of the data reported to UNESCO, and this, in turn, will definitely affect the quality of the outputs of UNESCO (e.g. statistics coverage, indicator calculations, database completeness).

The use of administrative data in the domain of education statistics has both its advantages and disadvantages. Some of the advantages mentioned by UNESCO-UIS (2008) are: system-wide (broad) coverage, annual basis compilation, cost efficiency (as countries usually collect the data for administration purposes), possibility of explicit link between student and resources (teachers, finances). Among the disadvantages, UNESCO-UIS (2008) mentions: lack of link to person or household socioeconomic status (either race or linguistic group), exclusion of non-formal education (outside the administration of school systems), and dependency on national statistical capacities to produce complete or accurate data. At the same time, related to the use of administrative data, Radermacher *et al.* (2009) comments that:

- The management of statistical register systems (production systems) is more complex than the management of statistical systems oriented around survey data (this complexity could become a challenge for nations with insufficient statistical capacities); and that

- Quality is a major challenge as data collections of administrative data are affected by different methods and definitions to those used on data collections based on survey data.

Certainly, there also seems to be a correspondence between the activities needed to create statistical records (and the related outputs) at the national level and the previously referred statistical programme that transforms national statistics into international comparable indicators (e.g. handling of multiple data sources, data validation, treatment of missing values, etc.). Indeed, these activities seem to be common in the domain of statistic production. Nevertheless, there are also important differences in production of statistics which depend on the type of survey (sample survey, census and administrative or register-based survey) used to collect the data. Some important differences relevant to the education international data collection are (Wallgren and Wallgren (2007 : 65):

- The success of sample surveys depends on the correct survey design and correct estimation of parameters and uncertainties. However, the success of censuses and register-based surveys depend on a "system-based thinking", in other words, their success depend on how well the different aspects of data production, edition and dissemination are conceived and coordinated. In the case of the education statistics collected by the UIS, the system-based thinking may enforce the inclusion of key elements in their data collection process, such as the conceptual framework for education statistics, the quality issues in data collection and the field work needed to help strengthening national statistical capacities in member states.
- Sample survey and censuses are owners of the data collection, including the production of their own questionnaires. Instead, registered-based surveys use the administrative records from multiple sources and owners. In the case of national education statistics, the ministry of education may have to oversee and validate the administrative records of enrolment and teachers that are reported by multiple sources: schools, institutes, universities, etc.
- Sample surveys must consider quality errors such as sampling errors and measurement errors (censuses would focus on measurement errors). However, registered-based surveys must definitely focus on different types of errors: lack of comparability between objective statistical variable and the administrative records available, timeliness (as administrative records may have slower processing times than sample survey), etc.

UIS' indicator development activities may also be seen as the (international) extension of the work done by the national organization that process and report outputs based on these types of data. In this regard, any international agency wishing to reliably compile national data must also pay attention to quality of its data sources in order to measure their impact on the quality of its indicator development activities and to assure the correct maintenance and sustainability of its data collection. Strategies about missing values or estimations as well as the current analysis on the stage of education in the world may vary depending on if the source of data is a sample survey (which may permit the development and inclusion of customized questions) or administrative data (which is more rigid with respect to the data available for collection but is expected to be more frequent in time).

2.2 Data quality and missing values

It is important to recognize at least three essential characteristics of reliable international education indicators: the conceptual and statistical framework for education statistics, data quality issues and the fieldwork required to build statistical capacity and sustain data collection at national levels. The present section will mainly focus on data quality issues and missing values. Although the relationships between missing values, the statistical framework and the activities of statistical building are also important, most of them could be considered as quality issues.

2.2.1 Data quality and its dimensions

As seen before, some important questions about education systems are usually responded through the analysis of high quality indicators.

In all cases, whether international education indicators are used for goal monitoring, for international comparison or for the study of certain education system elements, guaranteeing the adequate quality of data is essential in order to accurately inform policy makers or researchers (UNESCO-UIS, 2008). In this regard, the trust in the quality of the information produced by a statistical agency is critical for its survival (Brackstone, 1999) and therefore quality improvement must be a major part of its functions.

Although it is not within the scope of the current literature review to cover all aspects related to the quality of statistical information, some relevant concepts impacting UIS data collection will be presented next.

Quality has been defined in many ways. Hoyle (2009 : 24) listed the following definitions:

"- A degree of excellence (Oxford English Dictionary) – The meaning used by the general public.

- Freedom from deficiencies or defects (Juran) – The meaning used by those making a product or delivering a service.

- Conformity to requirements (Crosby) – The meaning used by those designing a product or a service or assessing conformity.

- Fitness for use (Juran) – The meaning used by those accepting a product or service.

- Fitness for purpose (Sales and Supply of Goods Act 1994) – The meaning used by those selling and purchasing goods.

- The degree to which a set of inherent characteristics fulfils requirements (ISO 9000:2005) – The meaning used by those managing or assessing the achievement of quality.

- Sustained satisfaction (Deming) – The meaning used by those in upper management using quality for competitive advantage."

Data quality or information quality related to the statistical outputs has been defined as "the fitness for use by clients" (Statistics Canada, 2002 : 2). In the case of UIS, its clients are: policy makers, governments, national statistical offices, general public, etc.

An interesting definition that take into account the fact that data are used for decision making is given by Karr *et al.* (2006 : 138): "*Data quality is the capability of data to be used effectively, economically and rapidly to inform and evaluate decisions*. Necessarily, DQ is multi-dimensional, going beyond record-level accuracy to include such factors as accessibility, relevance, timeliness, metadata, documentation, user capabilities and expectations, cost and context-specific domain knowledge."

This last definition introduces us to the approach statistical agencies – including UIS – use to make operative the concept of quality applied to statistical outputs: the definition of quality dimensions. Certainly, there is some degree of agreement in the literature about which the appropriate dimensions of data quality are and what experts consider the characteristics that a meaningful set of indicators must display. For example:

- OECD quality framework sets eight dimensions: relevance, accuracy, credibility, timeliness, punctuality, accessibility, interpretability, coherence (OECD, 2004).

- Statistics Canada considers six dimensions of quality for statistical outputs: relevance, accuracy, timeliness, accessibility, interpretability and coherence (Statistics Canada, 2002).
- For Rowe and Lievesley (2002), the quality elements of useful performance indicators are: validity, reliability, relevance to policy, potential for disaggregation, timeliness, coherence across different sources, clarity and transparency with respect to known limitations, accessibility and affordability, comparability through adherence to internationally agreed standards, consistency over time and location, and efficiency in the use of resources.
- For de Vries (2001) (based on United Nations sources), indicators should be assessed against criteria such as: policy-relevance, specificity, validity, reliability, sensibility, measurability, user-friendliness and cost effectiveness. De Vries summarizes this list in four desirable properties of an indicator: technical soundness, understandable, relevance and cost-effectiveness.
- For Batini and Scannapieco (2006), relevant quality dimensions related to data for information systems are (making reference to common areas where quality improvement is usually needed): accuracy, reliability, timeliness, completeness and consistency. In a more empirical point of view, these authors describe additional dimensions of quality, which they group in 4 categories: "(1) intrinsic data quality: believability, accuracy, objectivity and reputation; (2) contextual data quality: value-added, relevancy, timeliness, completeness and appropriate amount of data; (3) representational data quality: interpretability, ease of understanding, representational consistency, and concise representation; (4) accessibility data quality: accessibility and access security" (Batini and Scannapieco (2006 : 38).

There are many dimensions of data quality that are critical for national and international education statistics, such as relevance, accuracy and timeliness (UNESCO-UIS, 2008). Statistics Canada (2002 : 3) describes relevance, accuracy and timeliness as:

- "Relevance: The relevance of statistical information reflects the degree to which it meets the real needs of clients. It is concerned with whether the available information sheds light on the issues of most importance to users."
- "Accuracy: The accuracy of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure. It is usually characterized in terms of error in statistical estimates and is traditionally decomposed

into bias (systematic error) and variance (random error) components. It may also be described in terms of the major sources of error that potentially cause inaccuracy (e.g., coverage, sampling, nonresponse, response)."

- "Timeliness: The timeliness of statistical information refers to the delay between the reference point (or the end of the reference period) to which the information pertains, and the date on which the information becomes available. It is typically involved in a trade-off against accuracy. The timeliness of information will influence its relevance."

Quality dimensions can overlap and are definitely interrelated; in addition, there is not theoretical model that can bring all these dimensions into a single measure of quality (Statistics Canada, 2002). There is also a need to balance the sometimes conflicting activities needed to assure high data quality, including the multiple demands or expectations from data users (e.g. policy-makers, researchers and general public), while considering the related cost, the restrictions of limited resources, and even the burden imposed on national respondents. For example, taking more time to process education surveys may increase the accuracy and completeness of the data (e.g. dedicating more resources to find information from secondary sources, to make and validate estimations for missing values based on these secondary sources, or consulting with country officials to encourage the production of national estimates), but this would possibly have a negative impact in timeliness, as these actions, given the constricted schedule of national officials and the complexity of integrating different national data sources, demand considerable time and resources.

2.2.2 Completeness and missing values

The present research on missing values affects one particular dimension: completeness.

Completeness as a quality dimension has been generally defined as the degree or level to which the data are of sufficient breadth, deepness, and scope for the required functions or tasks (Batini and Scannapieco, 2006). In the case of UIS education statistics, a set education indicators must be "sufficiently complete" in order to adequately support the decisions of policy-makers (e.g. monitoring national and international goals, comparing national performance against competitors), the studies carried by researchers and the public debates.

To understand the sense of "sufficiently complete", first we must identify three possible types of completeness:

- Schema completeness. It refers to *"the degree to which concepts and their properties are not missing from the schema"* (Batini and Scannapieco, 2006 : 24). A "schema" is

the documentation describing the variables (and their respective metadata) included in a database.

- Column completeness. It refers to the "*measure of the missing values for a specific property or column in a table*" (Batini and Scannapieco, 2006 : 24). This is the most cited form of completeness, affecting data from all types of surveys, censuses and experiments.
- Population completeness. It refers to the "(evaluation of) *missing values with respect to a reference population*" (Batini and Scannapieco, 2006 : 24).

As we can see, completeness is intrinsically related to the problem of missing data – a phenomenon pervasive in empirical research, including social sciences (Allison, 2001; Graham, 2009). Indeed, survey methodologists have also differentiated between "unit non response" (the data collection fails for a sampled subject) – related to population completeness – and "item nonresponse" (only partial data available for the sample subject) – related to column completeness (Schafer and Graham, 2002).

In this regard, for the UIS education database or any subset of education indicators, "adequately complete" means that each education indicator from any given set must be clearly identified and described in the same database or in auxiliary documentation, that a given variable or indicator must not have a significant proportion of its values judged as missing (with respect to the estimated quantity of values needed to produce sound conclusions), and that the UNESCO respondents (target population is constituted by all UNESCO Member States) are all present in the international data collection.

Missingness (incomplete data) creates problems in scientific research because traditional statistical procedures are not designed to deal with missing data (Schafer and Graham, 2002). But correct handling of missing value is important in order to arrive to sound conclusions. As Schafer and Graham (2002 : 147) put it:

"Missingness is usually a nuisance, not the main focus of inquiry, but handling it in a principled manner raises conceptual difficulties and computational challenges. Lacking resources or even a theoretical framework, researchers, methodologists, and software developers resort to editing the data to lend an appearance of completeness. Unfortunately, ad hoc edits may do more harm than good, producing answers that are biased, inefficient (lacking in power), and unreliable."

Other problems with missingness are the resources and time needed to understand its nature, its impact on subsequent analysis and on the overall quality of the database and, based on

examination, to decide the adequate course of action. If corrective measures are needed, then the estimation and the validation of certain or all missing values with national officials add further burden to the data collection process.

Although no information about missing values on the UIS education database is available, their impact on analysis and conclusions are evident.

2.2.3 An example of missing values in monitoring education

This section presents an example of the effect that missing values has on relevant analyses. An extract of an analysis - carried by the UIS – that focuses on out-of-school children world-wide is presented below (UNESCO-UIS, 2011a : 40):

"Figure 18 [not reproduced here] compares the out-of- school rate of children of lower secondary school age in 1999 and 2009. In sub-Saharan Africa, the lower secondary out-of-school rate fell from 55% to 37% over this period, more than in any other region. Large reductions in the out-of-school rate were also observed in the Arab States (from 30% in 1999 to 16% in 2009). Other regions showing a substantial decrease during this period in the share of lower secondary school age children who are out of school are Central Asia (13% to 5%), East Asia and the Pacific (20% to 13%), and Latin America and the Caribbean (11% to 5%) [1]. However, progress was not universal; in Central and Eastern Europe, the percentage of out-of-school children increased from some 7% in 1999 to 11% in 2009. [Footnote 1: the regional average for East Asia and the Pacific is based on provisional UIS estimates.]"

The previous extracted analysis compares the performance related to the decrease (or increase) of the rate out-of-school children of lower secondary school age (indicator code: ROFSC-ISCED 2) among regions and between two specific years: 1999 and 2009. We can note that the analysis of South and West Asia's rate of out-of-school children is missing. The reason for this is that data for ROFSC is missing for South and West Asia for 1999 (UNESCO-UIS, 2011a).

We could imagine some of the issues an analyst must solve when dealing with missing values in his or her analysis:

(Note: the analysis of missing values is mostly made during the data validation following the data submission by country respondents. Given the characteristics and requirements of data users world-wide, the data collection as well as the work on data issues must be a continuous process, so any statistical data validation process must be able to cope with corrections on data and indicators at any time. Nevertheless, it is also true that, the sooner the missing data

treatment is done, the lesser the resources involved, and the higher the quality of the conclusions based on those datasets)

1) The analyst must look for the cause of the missing value of the regional indicator: Is it because there are not enough reporting countries in S.W. Asia in 1999? Is it because no reliable population data is available for S.W. Asia in 1999? The analyst must also search for others primary sources (in this case, number of out-of-school (OFS) children, instead of the rate of OFS children) or secondary sources (e.g. national reports). There are some corrective measures that can be proposed: using data from a year close to 1999 (1997 to 2001), estimation based on primary sources (extrapolation) or on secondary source, etc.

2) In this example, the (missing) value of indicator "ROFSC-ISCED 2" for South and West Asia (S.W. Asia) for 1999 was not estimated. The analyst must evaluate if the presented conclusions are being affected by not having any estimation of S.W. Asia's "ROFSC-ISCED 2" in 1999. In this case, important issues like the positive or negative evolution of ROFSC in S.W. Asia could not be verified.

3) Table 2 shows the regional (average) values of the rate and the number of OFS children of lower secondary age (ISCED 2) for 2009 that were used in the previous analysis. The double star code (**) qualifies data as UIS estimates while indicating that "the publishable data represent less than 60% of the relevant population" (UNESCO-UIS, 2011a : 87). Indeed, Table 2 is an example of the impact that missing data have on regional estimations - most of the indicators are qualified as UIS estimates (**). As we have seen before, it is probable that the input data (responses to the UNESCO education survey) did not include a significant proportion of countries or population compared to the number of countries or the total population in the specific region or that reported data for a significant quantity of countries were deemed as unreliable. In any case, given that each regional average is based on country data, we can also hint the intensive algorithmic, computational and human work that is needed to verify that each country and each regional average is correctly qualified, either as observed value, as a UIS estimation, or as missing data (not available or non publishable).

Table 2. Regional average for out-of-school children and the number of out-of-school children (lower secondary) - year 2009

Region averages	Out-of-school children of lower secondary school age				
	Out-of-school rate (%)			Number out of school	
	MF	M	F	MF (000)	% F
World	17**	16**	19**	71,608**	52**
Arab States	16**	12**	19**	3,507**	61**
Central and Eastern Europe	11**	10**	11**	2,089**	51**
Central Asia	5	4	6	377	57
East Asia and the Pacific	14**,-1	...	...	14,497**,-1	...
Latin America and the Caribbean	5**	5**	5**	1,948**	49**
North America and Western Europe	3	3**	3**	815	51**
South and West Asia	26**,-2	23**,-2	29**,-2	27,625**,-2	54**,-2
Sub -Saharan Africa	37**	33**	40**	21,637**	55**

Source: Data from UNESCO-UIS (2011 : 143)
Note: (**) UIS estimation; (...) missing value; ($^{-1}$,$^{-2}$) figures for 2008 and 2009 respectively.

According to the previous discussions, we can conclude that the best possible scenario for assuring high-quality low-cost data collection, data analysis and data dissemination is to have all Member States to submit complete and reliable data to the UIS. In this regard, data for all countries will increase the power of analysis and a complete questionnaire will reduce the time spend on analysis of missing values and in any corrective measure. Moreover, reliable data will reduce the quantity of data that is deemed as non publishable, which in turn will reduce the validation costs and speed processing times in the production of indicators and the respective analysis.

Note on the quality of statistic production

One of the main statistical outputs of UIS is its education database. It comprises more than 900 indicators and raw data points. Its elaboration involves the handling of electronic questionnaires, algorithms and methodologies for indicator calculation, data storage and dissemination in multiple formats - website and publications, and to multiple users – national statistical offices, ministries of education, etc. Due to the inherent complexity of the UIS data collection, it may seem that the previous reference to the possible data quality dimensions may not take full

account of the impact that the technological capabilities needed to maintain a reliable database or to manipulate adequately multiple sources of data have on overall quality (or either the impact of current UIS internal processes).

Batini and Scannapieco (2006) also discuss the different types of knowledge (input of information) present in an organization that have an impact in the methodology for treating data quality. We can find three groups of methodology inputs (Batini and Scannapieco, 2006 : 132): (1) the organizational knowledge, which includes mainly the business processes and its inherent quality, the organizational norms, rules and structures involved in the production of services, and the users requesting these organization outputs; (2) the technological knowledge, which includes all the databases and flow of information of the organization (collection of data), external sources of data, and; (3) the quality knowledge, which includes the dimensions of data quality described previously.

Given the nature of the UIS multiple-source data collection and its information or data-based products, the strengthening of its technological knowledge is crucial for assuring high quality data.

CHAPTER 3. Descriptive analysis of the UIS education database

3.1 Description of the UIS database of education statistics

3.1.1 Introduction

This section describes the UIS education database as well as a brief reference to the steps needed in the preparation of the data for the exploration and analysis of its missing values/completeness.

The current research is based on the freely and publicly available education database (UNESCO-UIS Data Center) downloaded in June 2011. There are three data releases per year (January, May, and October), and the list of variables may vary slightly among releases.

The UIS education database is a collection of 947 variables for 209 countries and territories. This statistical repository includes data since 1975.

The current research will be based on data from 1999 to 2009. The reason for choosing 1999 as start year is that the Institute of Statistics was created in 1999. Therefore, data from 1999 onwards can be considered to be consistently collected. As well, there is a delay of at least one year between the year of collection and the year of reference. For example, data for 2010 will be collected and compiled during 2011. Therefore, the assumption underlying the choice of 2009 as ending year is that data for this year have been collected and compiled during 2010.

Variables in the UIS education database can be classified in 2 types: raw data (468) and indicators (479). The main difference between raw data and indicators is that all raw data are directly collected from countries by a standardized measurement instrument (the annual UIS education survey), while indicators are produced from the raw data previously collected, as well as population data and some economic indicators provided by other United Nations' agencies.

Table 3. Distribution of variables in the UIS database by type

Type of variable	Number
Raw data	468
Indicators	479
Total	947

Indicators can be further divided in 2 types: country-level indicators and regional-level indicators. Country-level indicators are the statistics that describe aspects of the national education system. A regional level indicator is, in most cases, the average value of a given indicator for the countries of a certain region. In this regard, regional indicators do not add

information about countries' missing values. Therefore, the 145 regional indicators are withdrawn from analysis.

Most indicators and raw data are related to general characteristics of the educational system, such as enrolment, repeaters and teaching staff by age, by grade or by education level. Nevertheless, there are three groups of raw data variables that display a greater level of details than other groups:

- Enrolment in tertiary education by type of program (e.g. education, general education, humanities, social sciences, etc.) and by sex (total and female) (20 variables).
- Graduates in tertiary education by type of programme and by sex (20 variables).
- International students in tertiary education by country of origin (220 variables).

The analysis of missing values will not take into consideration these indicators.

Therefore, the number of variables that will be part of the exploratory analysis is 542, distributed as follows:

Table 4. Distribution of retained variables in the UIS database by type

Type of variable	Number
Raw data	196
Indicators	346
Total	542

Indicators and raw data can be also classified in concept-based groups and subgroups. This classification system will prove useful in the subsequent analysis as it facilitates the understanding of the analysis configurations and results without looking at the exact description of each indicator.

The summaries (Table 5 and Table 6) below contain the parents, subgroups, concepts, parameters and number of variables related to indicators and raw data. The information presented in these tables is not explicitly described by the UIS; it was prepared based on the parameters downloaded at the query stage.

In brief, for indicators, there are six parent groups: participation (114 indicators), entry (18), completion (21), progression (113), expenditure (50) and teacher (30); and for raw data, there

are six parent groups: participation (99), entry (4), completion (2), progression (40), teacher (34), and system (17).

Note that: by Sex = Total, male or female; by ISCED (education level) = ISCED 0 (pre-primary), ISCED 1 (primary), ISCED 2 (lower secondary), ISCED 3 (upper secondary), ISCED 4 (post secondary no tertiary), ISCED 5+6 (tertiary); GPI = gender parity index related to certain indicator.

Table 5. Indicator classification system: summary

Parent	Subgroup	Concept	Parameters	Number of variables
Participation	**Female participation**	Percentage of female students	by ISCED	**10**
	Over / under age	Over-age enrolment ratio	by Sex	**3**
		Under-age enrolment ratio	by Sex	**3**
	School age enrolment	Pupils of the official school age	ISCED 0,1,2-3 / by Sex	**9**
	Enrolment in tertiary	Number of students per 100 000 inhabitants	ISCED 5+6 / By Sex	**3**
	Gross enrolment ratio	Gross enrolment ratio	ISCED 1 to 6 / by Sex + GPI	**4**
		Gross enrolment ratio	ISCED 0,1,2,3, 5+6, 1+2; 1+2+3 / All programmes / by Sex + GPI	**28**
	Net enrolment rate	Adjusted net enrolment rate	ISCED 1 / by Sex + GPI	**4**
		Net enrolment rate	ISCED 0,1, 2+3 / by Sex + GPI	**12**
	School life expectancy	School life expectancy (years)	ISCED 1+2+3; 1+2+3+5+6; 5+6; 0 / + GPI	**15**
	Out-of-school children	Rate of primary school age children out of school	ISCED 0+1 / by Sex	**3**
		Rate of primary school age children out of school but in pre primary education	ISCED 1 / by Sex	**3**
	Student mobility indicators	International students %	Female	**1**
		Outbound mobility ratio %		**1**

Parent	Subgroup	Concept	Parameters	Number of variables
		Gross outbound enrolment ratio		**1**
		Inbound mobility rate		**1**
	Programme orientation	Technical/vocational enrolment ISC as % of total enrolment ISC	ISCED 2, ISCED 3, ISCED 2+3	**3**
	Distribution of tertiary students	Distribution of students (%)	ISCED 5A, 5B, 6	**3**
	Private education	Percentage of private enrolment	ISCED2- Gen, ISCED2-Tec, ISCED0, ISCED1, ISCED 2+3, ISCED 3-Gen, ISCED 3-Tec	**7**
Entry	**New entrants**	Percentage of new entrants to primary education with ECCE experience	ISCED1 / by Sex + GPI	**4**
	Intake to primary	Gross intake ratio	ISCED1 / by Sex + GPI	**4**
		Net intake rate	ISCED1 (theo.age; Over-age; Under-age) / by Sex + GPI	**10**
Completion	**Graduates ratios**	Expected gross primary graduation rate	ISCED1 / by Sex* + GPI	**4**
		Gross completion rate	ISCED 5A / by Sex + GPI	**4**
		Gross primary graduation rate	ISCED1 / by Sex + GPI	**4**
	Percentage tertiary graduates by programme	Percentage of female graduates	ISCED 5	**1**
	Proxy completion	Expected gross intake ratio to the last grade of primary	ISCED1 / by Sex + GPI	**4**
		Gross intake ratio to the last grade of primary	ISCED1 / by Sex + GPI	**4**
Progression	**Repetition rates**	Repetition rate	ISCED1 / by Grade1 (1 to 7) / by Sex	**21**
	Survival	Survival rate	ISCED1 / grade 4, 5, last grade / by Sex + GPI	**12**

Parent	Subgroup	Concept	Parameters	Number of variables
	Percentage of repeaters	Percentage of repeaters	ISCED1, ISCED 2 / by Grade (1 to 7, all grades; 1 to 8, all grades) / by Sex + GPI	**52**
	Transition	% Transition from ISCEDED 1 to ISCEDED 2	General programmes; by Sex + GPI	**4**
	School age	School age population	By ISCED (0,official entrance age, 1, 2, 3, 2+3, 4, 5+6); by Sex	**24**
Expenditure	**Public current expenditure**	Percentage distribution of public current expenditure by level	By ISCED (0, 1, 2, 3, 2+3, 4, 5+6, not allocated by level)	**8**
		Public current expenditure on education as % of total current government expenditure		**1**
		Public current expenditure on education as % of total public education expenditure		**1**
	Educational expenditure by nature / ISCEDED	Educational expenditure in ISCED as % total educational expenditure	by ISCED (0,1,2,3,4,2+3, 5+6, not allocated)	**8**
		Educational expenditure by nature of spending as a % of total educational expenditure on public institutions	ISCED 1+2+3+4, 5+6; Capital, total current exp., other current exp., salaries	**8**
	Percentage of GDP / GNP	Current expenditure on education as % GNI		**1**
		Public expenditure on education as % of ...	GDP, total government expenditure, GNI	**3**
		Public expenditure per pupil as a % of GDP per capita	By ISCED (all levels, ISCED 1, ISCED 2+3, ISCED 5+6)	**4**
		Total expenditure on educational institutions and administration as a % of GDP	International source / All levels	**1**
		Total expenditure on	By type of source	**15**

Parent	Subgroup	Concept	Parameters	Number of variables
		educational institutions and administration as a % of GDP	(private, public, all sources) / by ISCED (0,1,2+3+4, 5+6, all levels)	
Teacher	**Female teachers**	Percentage female teachers	BY ISCED (0,1, 2, 3, 2+3)	**5**
	Pupil-teacher ratio	Pupil-teacher ratio	By ISCED (0, 1, 2, 3, 2+3)	**5**
	Trained teacher	Percentage of trained teachers	By ISCED (0, 1, 2, 3, 2+3) / By Sex + GPI	**20**

Table 6. Raw data classification system: summary

Parent	Sub group	Concept	Parameters	Number of variables
Participation	**Enrolment in tertiary**	Enrolment	ISCED 5A, 5B, 6, ISCED5+6 / Public and private / full-part time / by Sex	**10**
	Enrolment in lower secondary	Enrolment	ISCED 2 / by Institution / by Programs / by Sex	**12**
	Enrolment in post-secondary non tertiary	Enrolment	ISCED 4 / by Institution / by Sex	**4**
	Enrolment in pre-primary	Enrolment	ISCED 0 / All programs / by Institution / by Sex	**4**
	Enrolment primary	Enrolment	ISCED 1 / by Institution / All programmes / by Sex	**4**
	Enrolment primary	Enrolment	ISCED 1 / by Grade / by Sex	**18**
	Enrolment in secondary by grade	Enrolment	ISCED 2+3 / by Grade / by Sex	**22**
	Enrolment in secondary	Enrolment	ISCED 2+3 / by Institution / by Programme / by Sex	**12**
	Enrolment in upper	Enrolment	ISCED 3 / by Institution	**12**

Parent	Sub group	Concept	Parameters	Number of variables
	secondary		/ by Programme / by Sex	
	Student mobility indicators	Students from a given country studying abroad		**1**
Entry	**New entrants**	New entrants to Grade 1	ISCED1 / by Sex	**2**
		New entrants who have attended some ECCE programmes.	ISCED1 / by Sex	**2**
Completion	**Tertiary graduates by programme**	Total graduates in all programmes. Tertiary	by Sex	**2**
Progression	**Repeaters in primary**	Repeaters	ISCED1 / by Grade (1 to 7, unspecified, all grades) / by Sex	**18**
	Repeaters in secondary	Repeaters	ISCED2 / by Grade (1 to 10, all grades) / by Sex	**22**
Teacher	**Teaching staff by ISCEDED**	Teaching staff	By ISCED (0,1, 5A, 5B, 5+6) / Public and private / Full-part time, all programmes/ By Sex	**10**
			By ISCED (2,3, 2+3) / Public and private / Full-part time / by type of prog (all, gen, tech.) / by Sex	**24**
System	**System**	Entrance age	By ISCEDED (0, 1, 2A, 3A, 4A)	**5**
		Duration	By ISCEDED (0, 1, 2A, 3A, 4A)	**5**
		Duration of compulsory		**1**
		Starting age of compulsory education		**1**
		Ending age of compulsory education		**1**
		Starting month of academic year		**1**

Parent	Sub group	Concept	Parameters	Number of variables
		Ending month of the academic year		1
		Starting year of the academic year		1
		Ending year of the academic year		1

3.1.2 Accessing UIS database of education statistics

The Institute of Statistics of UNESCO is responsible for the collection and compilation of internationally comparable education statistics. After internal processing and validation, this information is made available (dissemination phase) through an online data centre (UNESCO-UIS Data Center) and, as part of the UN statistical databases, through the UN online data repository (UNDATA Data sets)

As for terms of use, UNdata (2011) stipulates that: "All data and metadata provided on UNdata's website are available free of charge and may be copied freely, duplicated and further distributed provided that UNdata is cited as the reference". The UIS adheres to the terms of use of United Nations' statistical databases and does not issue a statement about its database's terms of use.

The UIS datacentre contains statistics from the fields of education, science and technology, culture and communications, and literacy (socio-economic and demographics data is also available, but it is not collected directly by UIS). The current research focuses on education statistics, which is one of the main domains of UIS work.

3.1.3 Data extraction

The education database used for the following analysis was downloaded on June 2011 from the UIS statistical tables (UNESCO-UIS Data Center). The online data centre has a web-based interface that allows access to either predefined statistical tables, or to execute personalized data queries. In order to download the complete education database, it is necessary to carry out a personalized data query (Customs tables' section).

To build a query, three parameters are required: country, year, and variable name. In the case of accessing the entire data base, one alternative is to build one 2-dimensional table – including all countries and all variables – for each year of study.

Note: cells in the UIS data tables can contain both values and qualifiers. In this regard, each cell result may be considered as a vector of 2 elements: values and qualifiers. The importance of qualifiers for the data preparation will be discussed next.

3.1.4 Metadata symbols and value symbols

There is one important aspect to take into consideration when manipulating the UIS education database: each entry (data point) consists of one value symbol or number and in certain cases, one qualifier (metadata). Apart from a numeric value of the data point (applicable when the data point's value is greater than zero), three additional symbols may be used to report a value:

Table 7. Value symbols

Value symbol	Meaning
... (3 consecutive dots)	No data available – missing value.
- (hyphen)	Magnitude nil or negligible.
. (single dot)	Not applicable.

At the same time, the value of a data point, if it exists, can be qualified as observed (no symbol added), as national estimation (*) or as UIS estimation (**).

Table 8. Metadata (qualifier) symbols

Metadata (qualifier) symbol	Meaning
** (Two stars)	UIS estimation
* (One start)	National estimation

There are two cases to consider when dealing with missing values:

1) A variable with a negligible value or considered as inapplicable to the context [for example, a variable capturing enrolment in primary grade 8 in a country without a grade 8 will marked as "not applicable" (.)] still brings information about the studied variable,

and must not be considered as missing value. Only values reported as data not available (...) are indicating missing values.

2) The qualification of a value as UIS estimation implies that this data point was not originally reported by the country's respondents and the displayed value is an estimation made by the institute (national estimations are estimations validated by national officials; as such, they could be considered as part of the response to the education survey). In this regard, data points qualified UIS estimations could be incorporated in the analysis of missing values.

The metadata symbols can be used to estimate the impact of the institute's estimations on the completeness of the database (or equivalently, on country or variable response rates). However, for the UIS data centre, these metadata qualifiers are not parameters that can be defined at the query stage; they have to be treated after the data tables are created. In other words, by default, all data tables from this data centre show UIS and national estimations in addition to the observed values. Moreover, qualifiers are not displayed in a parallel (metadata) table linked to the data tables, but they are incorporated directly as symbols indexed to the values in the resulting data tables.

Because plain file formats, such as CSV (comma separated values) or text, do not capture the metadata symbols, the most convenient format to work is Excel, which captures them as cell comments. When working in data preparation with Excel, it is necessary to develop a Visual Basic routine (Excel macro) to extract each cell comment and place it into tables similar to the original (country by variables), and then to determine if it is UIS estimation or national estimation.

3.1.5 Matrices of response

Any variable or vector representing information through numeric values, value symbols and metadata symbols can be converted into a binary variable representing the occurrence or not of a given condition. In this regard, the information conveyed in the original series of datasets will be transformed into matrices of binary data representing conditions related to the capacity of the database to display information about a country's education system. These will be called response matrices.

Case A: Matrix of response for submitted data and UIS estimations

Let define $Z_{y,c,v}$ as the data point for a given (y) year, (c) country and (v) variable that can be represented by numeric values and value symbols.

Then $X_{y,c,v} = 1$ if $Z_{y,c,v} \neq$ "..." (missing value),

and $X_{y,c,v} = 0$ otherwise; for all y, c, v.

X will be 1 when Z is a numeric value, negligible or not applicable (the metadata information is disregard) and 0 when Z is reported as missing. This condition describes the presence of any type of information about the national education systems as 1. It also implies that the response matrix confounds the presence of UIS estimations and data submitted by countries.

Case B: Matrix of response for submitted data only

Let define $M_{y,c,v}$ as the data point representing the metadata information related to y, c and v (e.g. national estimation, UIS estimation).

Then $X'_{y,c,v} = 0$ if $Z_{y,c,v} =$ "..." or $M_{y,c,v} =$ "UIS estimation",

and $X'_{y,c,v} = 1$ otherwise; for all y, c, v.

X' will be 0 when the Z is missing regardless of the metadata, or when Z is any value qualified as UIS estimation. X' will be 1 in all other cases. In reality, missing values are not qualified by metadata values, and the qualification of UIS estimation only affects numeric values.

For case A (submitted data and UIS estimations) the construction of the response matrix is direct as there is only one condition to observe in the dataset associated with Z. For case B (only submitted data), there are two conditions to observe in two different datasets: values (related to Z) and metadata (related to M). The metadata table must be previously obtained by extracting all qualifiers from the original UIS data table.

Therefore, a 2-dimensional response matrix (evaluating X on $C = 209$ countries and $V = 542$ variables) is obtained for each year ($Y =$ 1999 to 2009) and for each of the referred case (A and B).

3.2 Descriptive analysis of country response rates

This section presents a descriptive analysis of the country response rates for case A and B. The analysis of response rates by country gives a quick glance at the behaviour of countries when reporting educational information, in other words, at the general statistical capacity of the country to respond to the UIS questionnaire.

As seen before, each response matrix contains data for 209 countries and 542 variables. There are eleven response matrices, corresponding to years 1999 to 2009.

A country response rate can be defined as the sum of responses X (or X') for all variables for a given country and year divided by 542.

Country response rate $(y, c) = \left(\sum_{v=1}^{542} X_{y,c,v}\right)/542$ for a given country c, for a given year y.

Country response rates vary between 0 (the country was not able to submit any national statistic and no indicators were calculated) to 1 (country submitted all questionnaire data and all indicators were calculated).

3.2.1 Quantitative description of country response rates - case A

Table 9 shows the basic statistics for country response rates by year (submitted data and UIS estimations). It can be seen that the average country response rates per year has been stable across time, varying from 0.546 in 1999 to around 0.59 in average between 2001 to 2007. Nevertheless, for 2008, the average country response rate seems to slightly decrease (0.57). Moreover, for 2009, the average country response rate decreased 20% with respect to 2008 (from 0.577 to 0.461). This significant decrease could be explained by different reasons: a considerable number of countries reporting education data that are 2 years behind the calendar year, delays in the submission of data or in the institute's data processing.

Table 9. Basic Statistics - case A (submitted data and UIS estimations)

Country response rate: Basis Statistics					
Year	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1999	209	0.546	0.228	0.024	0.897
2000	209	0.566	0.222	0.024	0.91
2001	209	0.583	0.23	0.031	0.924
2002	209	0.595	0.234	0.031	0.956
2003	209	0.594	0.241	0.052	0.923
2004	209	0.604	0.252	0.065	0.983
2005	209	0.598	0.247	0.031	0.991
2006	209	0.561	0.271	0.006	0.983
2007	209	0.594	0.248	0.07	0.991
2008	209	0.577	0.257	0.065	0.991
2009	209	0.461	0.278	0	0.932

The boxplot in Figure 1 compares country response rates across years. It can be seen that the average country response rate decreased in 2009 compared to previous years as well as a slight decrease in the average country response rate in 2006 (0.56) compared to that of 2005 (0.60) and 2007 (0.59). The relative larger spacing of the bottom of the boxes (the 25th percentile) to the average compared to the spacing of the average to the top of the boxes suggests a negative skewed distribution. This is better appreciated in Figure 2 (histograms of country response rate).

Figure 1. Box plot for country response rate - case A (submitted data and UIS estimations)

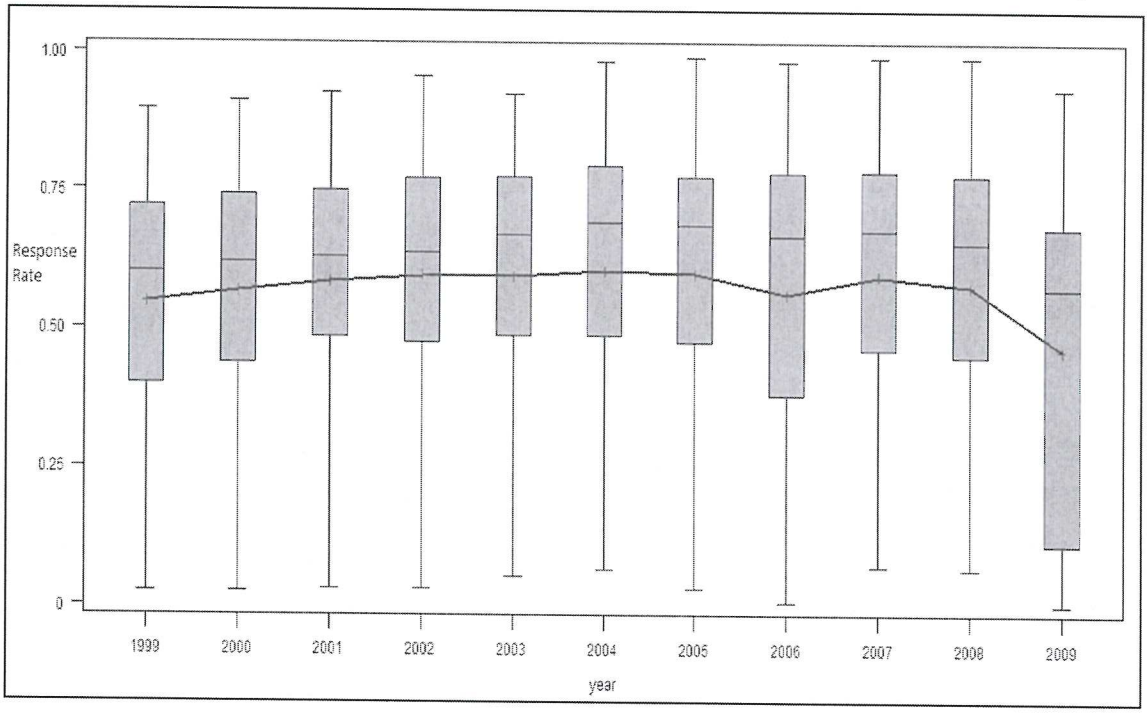


Table 10 shows the distribution of country response rates by year. The first column shows the bin (range) of response rates, and the table's values indicate the proportion of countries (from the total of 209) that fall within the given response rate range by year.

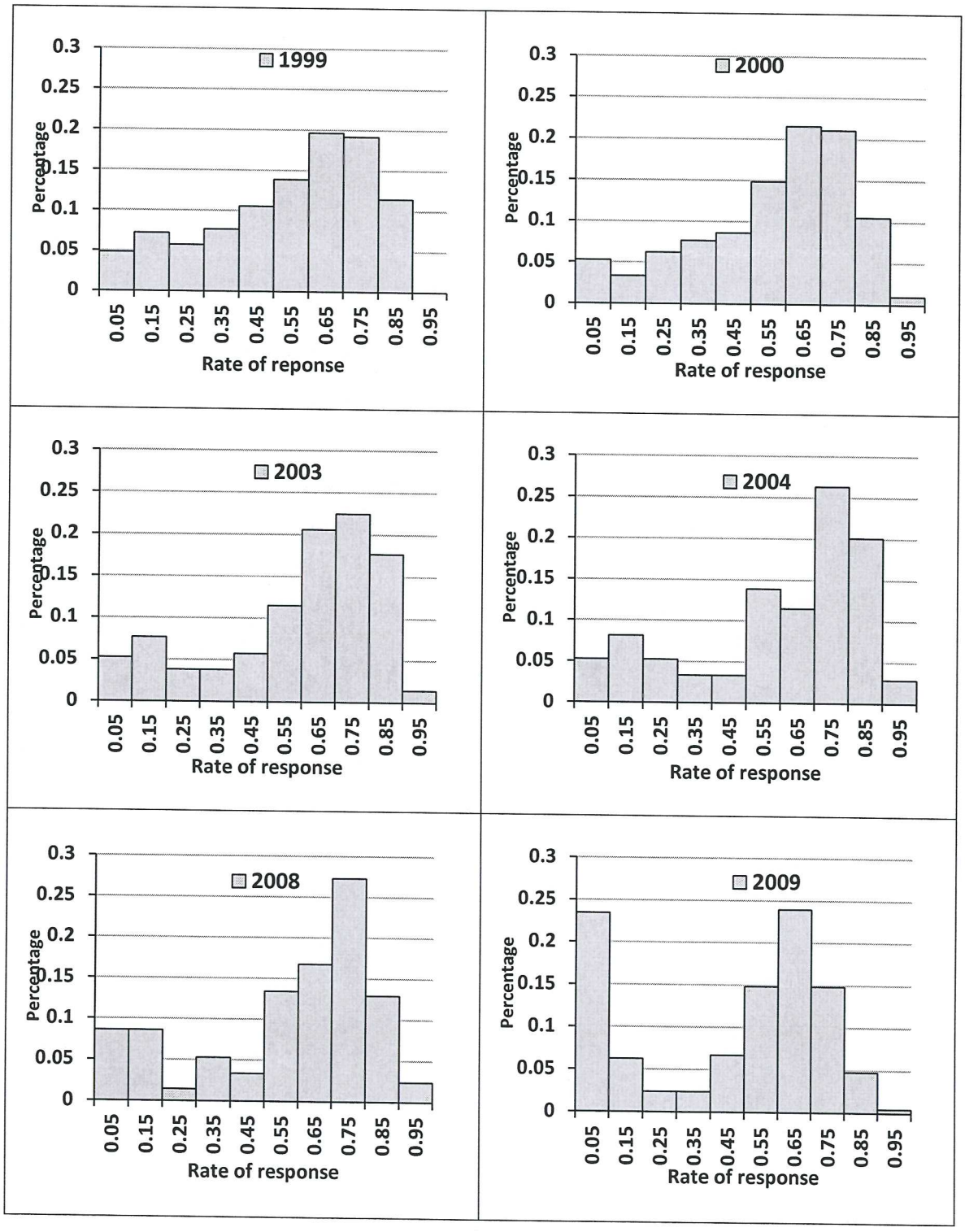
It can be noted that, across years, between 60% and 72% of countries fall within the response range of 0.5 and 0.9. The most frequent response range, from 1999 to 2008, is 0.7–0.8, with an average of 24% of countries falling within. For 2009, the most frequent range is 0.6–0.7, with 24% of countries, indicating a possible problem with the data submission. As well as with the average response rate, the distribution of the proportion of countries by response rate ranges and year seems stable through time.

Table 10. Distribution of country response rates – proportion of countries by response rate ranges – case A (submitted data and UIS estimations)

Response rate (bin)	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
0-0.1	0.05	0.05	0.05	0.05	0.05	0.05	0.07	0.11	0.07	0.09	0.23
0.1-0.2	0.07	0.03	0.04	0.04	0.08	0.08	0.06	0.08	0.06	0.09	0.06
0.2-0.3	0.06	0.06	0.06	0.05	0.04	0.05	0.05	0.03	0.03	0.01	0.02
0.3-0.4	0.08	0.08	0.07	0.06	0.04	0.03	0.04	0.07	0.06	0.05	0.02
0.4-0.5	0.11	0.09	0.06	0.08	0.06	0.03	0.05	0.04	0.07	0.03	0.07
0.5-0.6	0.14	0.15	0.16	0.16	0.11	0.14	0.11	0.12	0.10	0.13	0.15
0.6-0.7	0.20	0.22	0.16	0.14	0.21	0.11	0.15	0.13	0.15	0.17	0.24
0.7-0.8	0.19	0.21	0.23	0.21	0.22	0.26	0.28	0.25	0.26	0.27	0.15
0.8-0.9	0.11	0.11	0.15	0.14	0.18	0.20	0.17	0.15	0.17	0.13	0.05
0.9-1	0.00	0.01	0.01	0.06	0.01	0.03	0.02	0.02	0.03	0.02	0.00
0.5-0.9	0.64	0.68	0.71	0.65	0.72	0.72	0.71	0.66	0.68	0.70	0.58

The following histograms of country response rates (Figure 2) illustrate the previous observations with respect to the most frequent bins of response rates (0.5–0.9). Indeed, from 1999 to 2008, the pattern of distribution of country response rates seems to be similar (1 peak between 0.6 and 0.8, lower relative frequencies in lower response ranges). Nevertheless, for 2009, the distribution seems to have shifted to the left, while the proportion of countries in the response bin of 0.0–0.1 has increased significantly (0.23 in 2009 versus the average of 0.06 from 1999 to 2008).

Figure 2. Histograms of country response rate - case A (submitted data and UIS estimations) - selected years. Y-axis represents the relative frequency (percentage) and X-axis the bin of rate of responses.



3.2.3 Quantitative description of country response rates – case B

This section presents a quantitative description based on data originally submitted by the country or qualified as national estimations but do not include UIS estimations (case B). A comparison between case A (submitted data and UIS estimations) and case B (submitted data only) allows us to assess the effect of UIS estimations on the average country response rates.

Table 11 shows the basic statistics of country response rate – case B – and Table 12 shows the comparison of the average country response rate per year for case A and case B. From 1999 to 2005, UIS estimations counted for an average increase of 6.6 percentage points in the average country response rate, which is equivalent to the addition of 7540 responses (data points) to the data base each year. Nevertheless, from 2006 to 2009, the production of UIS estimations has decreased significantly, adding in average just 2.2 percentage points to the average country response rate during these years (see Figure 3).

Table 11. Basic Statistics - case B (submitted data only)

Country response rate: Basis Statistics					
year	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1999	209	0.487	0.224	0.024	0.88
2000	209	0.497	0.221	0.024	0.895
2001	209	0.509	0.231	0.031	0.902
2002	209	0.527	0.241	0.031	0.956
2003	209	0.509	0.25	0.05	0.886
2004	209	0.548	0.252	0.065	0.945
2005	209	0.543	0.249	0	0.983
2006	209	0.532	0.267	0.006	0.983
2007	209	0.568	0.251	0.031	0.991
2008	209	0.557	0.256	0.039	0.991
2009	209	0.447	0.271	0	0.902

Table 12. Comparison of country response rate - case A and case B

	Response rate		Difference	Number of UIS estimations
	Submitted and UIS estimations (A)	Submitted data only (B)		
year	Mean	Mean	A-B	
1999	0.546	0.487	0.059	6683
2000	0.566	0.497	0.069	7816
2001	0.583	0.509	0.074	8383
2002	0.595	0.527	0.068	7703
2003	0.594	0.509	0.085	9629
2004	0.604	0.548	0.056	6344
2005	0.598	0.543	0.055	6230
2006	0.561	0.532	0.029	3285
2007	0.594	0.568	0.026	2945
2008	0.577	0.557	0.02	2266
2009	0.461	0.447	0.014	1586

Figure 3. Evolution of the production of UIS estimations

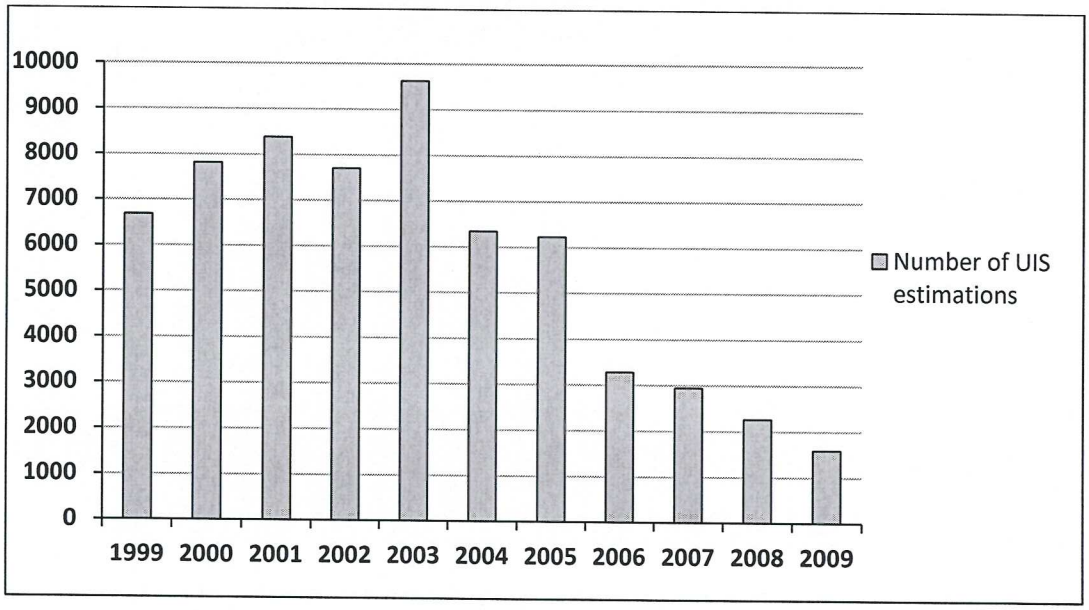
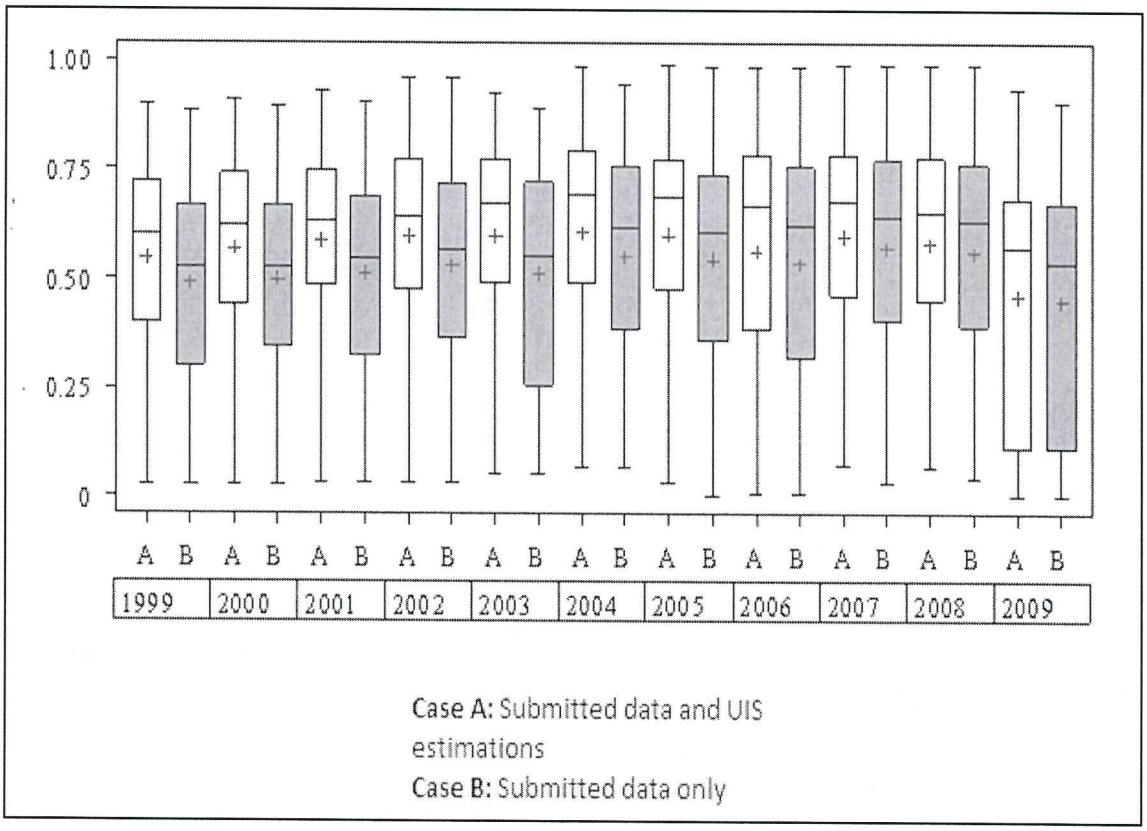


Figure 4 illustrates the effect of inclusion of UIS estimations in the average country response rate from 1999 to 2009. As mentioned before, UIS estimations production has declined since

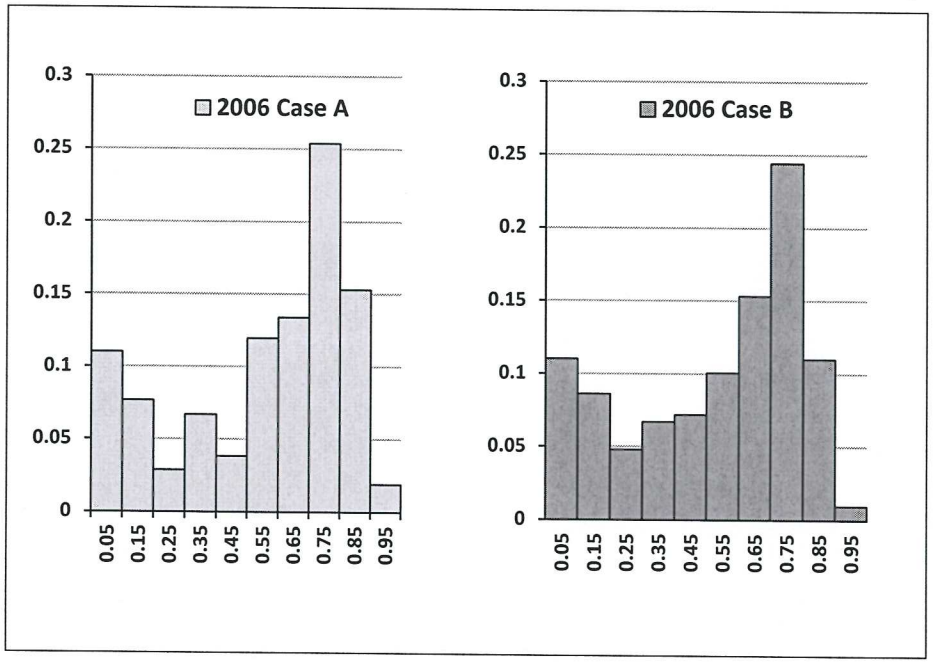
2006, making the average country response rate between case A and B almost identical in recent years (2008 and 2009).

Figure 4. Boxplots - Comparison of country response rate - case A and case B



Figures 5 shows the distribution of country response rates for submitted data and UIS estimations (case A) and submitted data only (case B) for year 2006. It can be noted that the inclusion of UIS estimations increases the proportion of countries falling in range of response of 0.7 and 0.9, and decreases the proportion of countries falling in lower ranges of response rates.

Figure 5. Comparison of the distribution of country response rates - case A and B - year 2006. Y-axis represents the relative frequency (percentage) and X-axis the bins of rate of response.



3.2.4 Ranking of country response rates – Top 10 and bottom 10 countries per year

The following rankings of the 10 countries with the highest and lowest response rates from 1999 to 2009 lets us to recognize a certain degree of stability in the countries belonging to these 2 groups. We notice that countries with different economic and social characteristic can be in the same top or bottom group. This descriptive analysis is based on data from submissions and UIS estimations (case A).

Table 13 shows the lists, per year, of countries in the top 10 of response rates. There are a number of countries that continuously perform well across years, being frequently in the top 10 list: Cuba (10 times from 1999 to 2009), Kyrgyzstan (8), Lao P. D. R. (8), R. Korea (7), Eritrea (6), Azerbaijan (5), Panama (5), El Salvador (4), Mexico (4), Mongolia (4), Niger (4). Other countries, like Venezuela, Mali, Serbia, Madagascar or Cambodia, have appeared at least once in these lists, demonstrating the potential of these national statistics systems of submitting complete education data (the statistical capacity may improve or deteriorate depending on economic or political factors). The range of the top 10 country response rates from 1999 to 2009 varies from 0.81 to 0.99, and the average is 0.90. No country shows a response rate of 1. Regarding the economic differences between countries in the top 10, as per the World Bank, the GDP per capita (current US$) for Cuba was 5 565 in 2008, for Kyrgyzstan, 881 in 2009, for Lao P.D.R., 997 in 2009 and for Republic of Korea, 17 110 in 2009 (source: World Bank Data

Catalogue). The diversity of the economic backgrounds of these countries may indicate that the improvement of the education statistical capacity in a country may depend, in addition to monetary investment, on a variety of factors such as political cooperation, stability of the institutions administrating education, internal demand for education indicators, etc.

Table 13. List of the 10 countries with the highest response rates by year – case A (submitted data and UIS estimations)

Year	10 countries with the highest response rates (response rate in parenthesis)
2009	Niger (0.93), Mali (0.89), Antigua & B. (0.88), Colombia (0.87), Cuba (0.87), Guyana (0.87), Moldova (0.83), Burkina Faso (0.83), United Arab E. (0.82), Central African R. (0.81)
2008	Cuba (0.99), Panama (0.94), Mali (0.94), Niger (0.93), Moldova (0.92), Kyrgyzstan (0.89), Lebanon (0.89), Venezuela (0.88), El Salvador (0.88), Serbia (0.87)
2007	Cuba (0.99), El Salvador (0.95), Moldova (0.92), Niger (0.91), R. Korea (0.9), Kyrgyzstan (0.9), Panama (0.89), Madagascar (0.89), Cambodia (0.89), Romania (0.88)
2006	Cuba (0.98), El Salvador (0.97), Lao P. D. R. (0.91), R. Korea (0.9), Azerbaijan (0.89), Kyrgyzstan (0.89), Niger (0.89), Cyprus (0.88), Belarus (0.88), Mexico (0.88)
2005	Cuba (0.99), Lao P. D. R. (0.93), Azerbaijan (0.9), El Salvador (0.9), R. Korea (0.9), Kyrgyzstan (0.89), Philippines (0.88), Belarus (0.88), Brunei D. (0.88), Mexico (0.88)
2004	Cuba (0.98), Lao P. D. R. (0.95), Panama (0.94), Eritrea (0.94), Mongolia (0.9), R. Korea (0.9), Kyrgyzstan (0.89), Morocco (0.89), Mexico (0.88), Colombia (0.88)
2003	Seychelles (0.92), Cuba (0.9), R. Korea (0.9), Azerbaijan (0.89), Mexico (0.89), Kuwait (0.89), Bulgaria (0.89), Lao P. D. R. (0.89), Cyprus (0.88), Eritrea (0.88)
2002	Cuba (0.96), Lao P. D. R. (0.93), Seychelles (0.92), Mongolia (0.92), Panama (0.91), Bulgaria (0.91), Eritrea (0.91), Croatia (0.91), Aruba (0.9), R. Korea (0.9)
2001	Panama (0.92), Cuba (0.92), Lao P. D. R. (0.9), Cambodia (0.9), Bulgaria (0.89), Aruba (0.89), Trinidad & T. (0.88), Azerbaijan (0.87), Eritrea (0.87), Kyrgyzstan (0.87)
2000	Eritrea (0.91), Lao P. D. R. (0.9), Mongolia (0.89), Bhutan (0.87), Morocco (0.86), United Arab E. (0.86), Azerbaijan (0.86), Samoa (0.85), Kyrgyzstan (0.85), Croatia (0.85)
1999	R. Korea (0.9), Eritrea (0.89), Mongolia (0.88), Morocco (0.88), Lao P. D. R. (0.88), Croatia (0.87), Cuba (0.86), Kyrgyzstan (0.86), Latvia (0.85), Trinidad & T. (0.85)

Table 14 shows the lists, from 1999 to 2009, of countries at the bottom 10 of response rates. As in the previous case, there are a number of countries that frequent these positions: Puerto Rico (10 times), Haiti (9), Bosnia and Herzegovina (8), Turkmenistan (8), San Marino (7), Liberia (6), Lebanon (5), Montenegro (4), Liechtenstein (4), Democratic People's Republic of Korea (4), Singapore (4) and Guinea-Bissau (4). The range of the bottom 10 country response rates from 1999 to 2009 varies from 0 to 0.09, and the average is 0.07. The GDP per capita (current US$) for Haiti was 657 in 2009, for Bosnia and Herzegovina, 4 523 in 2009, for Turkmenistan, 3 710

in 2009, for San Marino, 60 895 in 2008, and for Singapore, 36 758 in 2009 (source: World Bank Data Catalogue). In this regard, countries in the bottom positions also display a wide range of GDP per capita levels.

Table 14. List of the 10 countries with the lowest response rates by year - case A

Year	10 countries with the lowest response rates (response rate in parenthesis)
2009	Ecuador (0), El Salvador (0.02), Jamaica (0.02), Puerto Rico (0.02), Bahamas (0.02), Honduras (0.02), Anguilla (0.03), Turks & C. (0.03), Jordan (0.06), Yemen (0.06).
2008	Somalia (0.06), Papua New Guinea (0.07), Guinea-Bissau (0.07), Libyan Arab Jamahiriya (0.08), Turkmenistan (0.08), D. P. R. Korea (0.08), Micronesia (0.08), Palau (0.08), Tonga (0.08), Haiti (0.08).
2007	San Marino (0.07), Albania (0.07), Haiti (0.08), Puerto Rico (0.08), Turks & C. (0.08), Zimbabwe (0.08), Guinea-Bissau (0.08), Libyan A. J. (0.08), Turkmenistan (0.08), Netherlands Antilles (0.09).
2006	Iraq (0.01), San Marino (0.03), Singapore (0.04), Bosnia & H. (0.06), Albania (0.07), Timor-Leste (0.08), Puerto Rico (0.08), Trinidad & T. (0.08), Turks & C. (0.08), Liberia (0.08).
2005	San Marino (0.03), Iraq (0.04), Bosnia & H. (0.06), Singapore (0.07), Albania (0.07), Haiti (0.08), Puerto Rico (0.08), Liberia (0.08), Netherlands Antilles (0.09), Turkmenistan (0.09).
2004	Bosnia & H. (0.06), Haiti (0.08), Puerto Rico (0.08), Guinea-Bissau (0.08), Lebanon (0.08), Netherlands Antilles (0.08), Liberia (0.09), Turkmenistan (0.09), D. P. R. Korea (0.09), Singapore (0.1).
2003	San Marino (0.05), Bosnia & H. (0.06), Lebanon (0.07), Haiti (0.08), Puerto Rico (0.08), Canada (0.08), Guinea-Bissau (0.08), Micronesia (0.09), Liberia (0.09), Bhutan (0.09).
2002	San Marino (0.03), Montenegro (0.06), Bosnia & H. (0.06), Liechtenstein (0.07), Haiti (0.08), Puerto Rico (0.08), Liberia (0.09), Lebanon (0.09), D. P. R. Korea (0.09), Turkmenistan (0.1).
2001	Liechtenstein (0.03), San Marino (0.03), Lebanon (0.06), Montenegro (0.06), Bosnia & H. (0.06), Haiti (0.08), Puerto Rico (0.08), Turkmenistan (0.08), Chile (0.08), Liberia (0.09).
2000	Montenegro (0.02), Liechtenstein (0.03), Bosnia & H. (0.06), Andorra (0.07), Lebanon (0.07), Timor-Leste (0.07), Turkmenistan (0.08), Puerto Rico (0.08), Suriname (0.08), Haiti (0.09).
1999	Montenegro (0.02), Liechtenstein (0.03), Bosnia & H. (0.06), Timor-Leste (0.07), Andorra (0.07), San Marino (0.07), Turkmenistan (0.08), Haiti (0.08), Puerto Rico (0.08), D. P. R. Korea (0.09).

3.3 Descriptive analysis of variable response rates

We now proceed to present a descriptive analysis of the response rates per variable for case A, hence having a glance at which variables are the most available worldwide and which are more difficult to respond.

A variable response rate can be defined as the sum of responses X (or X') for all countries, for a given variable and year, divided by 209.

Variable response rate $(y, v) = \left(\sum_{c=1}^{209} X_{y,c,v}\right)/209$ for a given variable v, for a given year y.

Variable response rates vary between 0 (no country responded to the variable) to 1 (the variable was responded or calculated for all countries).

3.3.1 Quantitative description of variable response rates – case A

Table 15 shows the distribution of response rates per variable by year with data for case A (submitted data and UIS estimations). The first column shows the bin (range) of response rates and the values of the table indicate the proportion of variables (from the total of 542) that fall within the given response rate range by year.

There are some differences in the distribution of variable response rates compared to the distribution of country response rates. For example, across years, between 65% and 71% of variables falls within the response range of 0.4 and 0.8. Nevertheless, the bulk of country response rates (60% to 72%) are found within the response range of 0.5 and 0.9. In other words, countries have better performance in response rates than variables (more countries, proportionally speaking, have more information along variables than variables have information along countries).

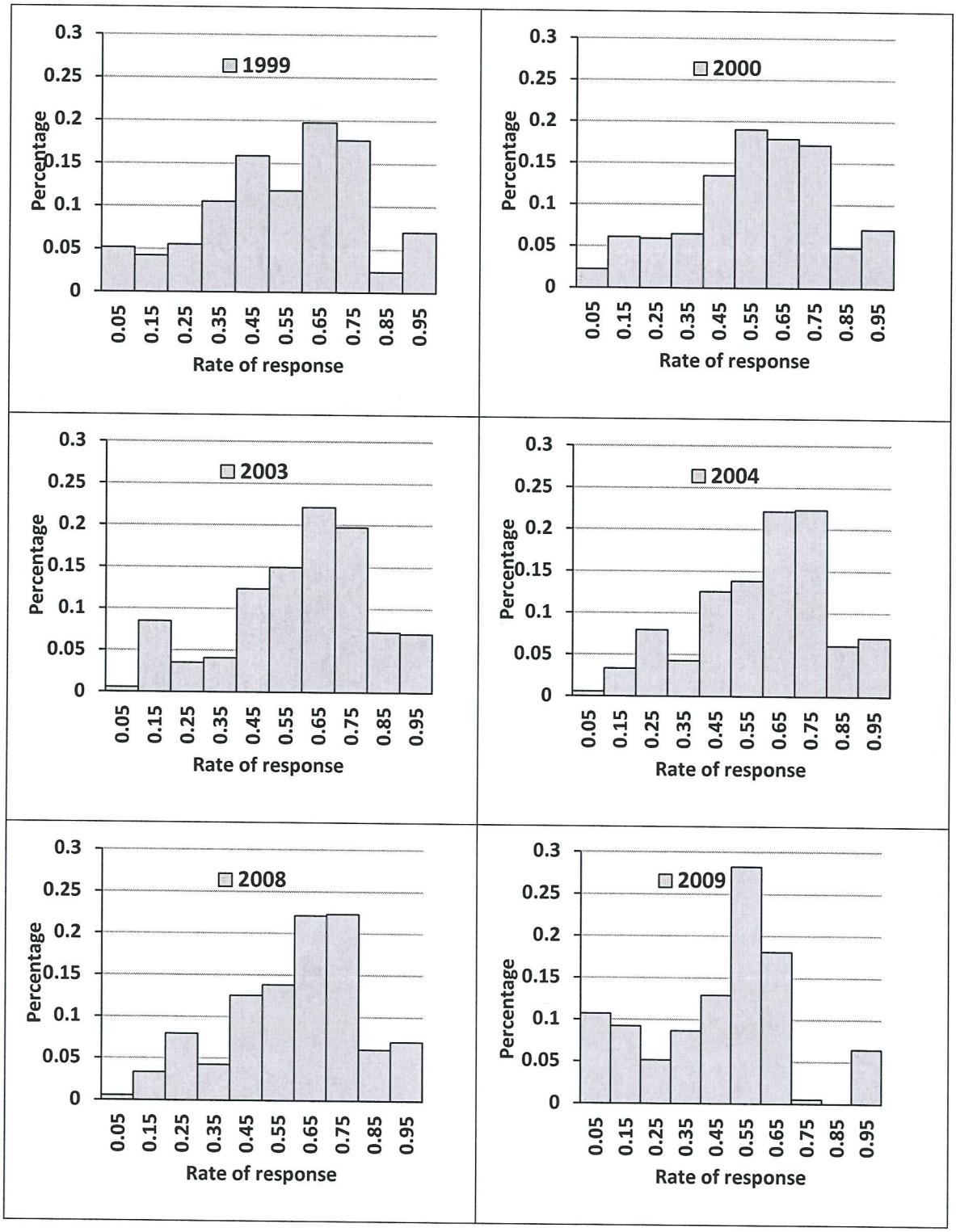
Regarding the variable response rates, the most frequent response range is 0.6–0.7, with an average of 21% of variables falling within this response range from 1999 to 2008. For 2009, the most frequent response range is 0.5–0.6 with 28% of variables, and the response range of 0.7–0.8 is 0, while the average of variables falling within this range is 19%. These indicate a possible problem with the data submission in 2009. Similarly to the country response rates, the distribution of the proportion of variables by response rate ranges and year seems stable across time. These facts can also be appreciated in Figure 6.

Table 15. Distribution of response rates per variable - proportion of variables by response rate ranges - case A

Response rate (bin)	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
0-0.1	0.05	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.11
0.1-0.2	0.04	0.06	0.04	0.05	0.08	0.03	0.06	0.07	0.04	0.04	0.09
0.2-0.3	0.06	0.06	0.06	0.04	0.04	0.08	0.04	0.05	0.06	0.06	0.05
0.3-0.4	0.11	0.06	0.08	0.06	0.04	0.04	0.07	0.10	0.08	0.09	0.09
0.4-0.5	0.16	0.13	0.11	0.14	0.12	0.13	0.11	0.14	0.12	0.15	0.13
0.5-0.6	0.12	0.19	0.18	0.16	0.15	0.14	0.14	0.16	0.18	0.14	0.28
0.6-0.7	0.20	0.18	0.19	0.22	0.22	0.22	0.21	0.24	0.17	0.24	0.18
0.7-0.8	0.18	0.17	0.20	0.16	0.20	0.22	0.21	0.15	0.22	0.18	0.01
0.8-0.9	0.02	0.05	0.05	0.08	0.07	0.06	0.07	0.02	0.06	0.01	0.00
0.9-1	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.06

0.4-0.8	0.65	0.68	0.68	0.68	0.69	0.71	0.68	0.69	0.69	0.71	0.60

Figure 6. Histograms of response rate per variable - case A (submitted data and UIS estimations) - selected years. The Y-axis represents the relative frequency (percentage) and X-axis the bins of rate of response.



3.3.2 Ranking of variable response rates – Top 10 and bottom 10 variables per year

The following rankings of the 10 variables with highest (top 10) and lowest (bottom 10) response rates from 1999 to 2009 let us recognize certain degree of stability in the variables belonging to these 2 groups. This descriptive analysis is based on data from country submissions and UIS estimations (case A). Variables describing some relatively permanent characteristic of national education systems, like entrance age by education level (e.g. primary, upper and lower secondary, etc.), duration by education levels, or variables that strongly depend on external sources, such as school age population by education level (which depend on entrance age and population data from the United Nations Population Division) have been excluded from the present analysis. In particular, due to their descriptive nature, these excluded variables exhibit the largest response rates each year.

Table 16 shows the most frequent variables at the top 10 response rates from 1999 to 2009 ("frequency" is defined as the number of years, from 1999 to 2009, that a given variable can be found among the top 10 or bottom 10 in the case of Table 18) and Table 17 shows the lists of top 10 variables per year. As in the case of countries, there are variables that perform well across years, among them: enrolment in primary (total and female), gross enrolment ratio (total and female), percentage of repeaters in primary (total and female), etc. It can be noted that most of the variables with the highest response rates are statistics (indicators and raw data) that measure participation or progression related to primary education. Without taking into consideration 2009, the range of top 10 variable response rates varies consistently between 0.80 and 0.89. For 2009, the range of response rates varies from 0.67 to 0.70.

Table 16. Variables at the top 10 response rate from 1999 to 2009 - case A (submitted data and UIS estimations)

Concept	Parent	Subgroup	Code	Frequency
Percentage of female students. Primary	Participation	Female participation	PFSI1	11
Enrolment in primary. Public and private. All programmes. Total	Participation	Enrolment primary	E20062	11
Enrolment in primary. All grades. Total	Participation	Enrolment primary	E21423	11
Enrolment in primary. Public and private. All programmes. Female	Participation	Enrolment primary	E20063	10
Enrolment in primary. All grades. Female	Participation	Enrolment primary	E21447	10

Concept	Parent	Subgroup	Code	Frequency
Technical/vocational enrolment in ISCEDED 2 as % of total enrolment in ISCEDED 2	Participation	Programme orientation	TVLSP	9
Gross enrolment ratio. Primary. Total	Participation	Gross enrolment ratio	GERFT	9
Gross enrolment ratio. Primary. Female	Participation	Gross enrolment ratio	GERFF	5
Percentage of repeaters in primary. Grade 7. Female	Progression	Percentage of repeaters	PRFF7	4
Percentage of repeaters in primary. Grade 7. Male	Progression	Percentage of repeaters	PRFM7	4
Percentage of repeaters in primary. Grade 7. Total	Progression	Percentage of repeaters	PRFT7	4
Enrolment in lower secondary. Public and private. General programmes. Total	Participation	Enrolment in lower secondary	E20066	4

Table 17. List of top 10 variable response rates by year - case A

Year	Top 10 variables (response rate in parenthesis)
2009	PFSI1 (0.7), E20062 (0.7), E20063 (0.7), E21423 (0.7), E21447 (0.7), E20066 (0.68), E20067 (0.68), E20084 (0.67), E20085 (0.67), E21323 (0.67).
2008	E20062 (0.84), E21423 (0.83), PFSI1 (0.83), E20063 (0.83), E21447 (0.82), E20066 (0.81), E20067 (0.81), GERFT (0.8), E21323 (0.8), E21347 (0.8).
2007	PRFF7 (0.88), PRFM7 (0.88), PRFT7 (0.88), E20062 (0.87), PFSI1 (0.86), E20063 (0.86), E21423 (0.86), E21447 (0.86), GERFT (0.84), TVLSP (0.84).
2006	PRFF7 (0.83), PRFM7 (0.83), PRFT7 (0.83), RR7FF (0.82), RR7FM (0.82), RR7FT (0.82), TVLSP (0.81), E20062 (0.8), E21423 (0.8), PFSI1 (0.8).
2005	E20062 (0.88), E21423 (0.88), PFSI1 (0.87), E20063 (0.87), E21447 (0.87), PRFF7 (0.86), PRFM7 (0.86), PRFT7 (0.86), GERFT (0.85), TVLSP (0.85).
2004	E20062 (0.87), E21423 (0.87), PFSI1 (0.85), E20063 (0.85), E21447 (0.85), TVLSP (0.85), GERFT (0.84), PRFF7 (0.84), PRFM7 (0.84), PRFT7 (0.84).
2003	E20062 (0.86), E21423 (0.86), PFSI1 (0.86), E20063 (0.86), E21447 (0.86), TVLSP (0.85), GERFT (0.84), GERFF (0.84), GERFM (0.84), GPGEP (0.84).
2002	E20062 (0.89), E21423 (0.89), TVLSP (0.88), PFSI1 (0.88), E20063 (0.88), E21447 (0.88), GERFT (0.88), GERFF (0.87), GERFM (0.87), GPGEP (0.87).
2001	E20062 (0.89), E21423 (0.88), TVLSP (0.88), PFSI1 (0.87), E20063 (0.87), E21447 (0.87), GERFT (0.87), GERFF (0.85), GERFM (0.85), GPGEP (0.85).
2000	TVLSP (0.89), E20062 (0.88), E21423 (0.88), PFSI1 (0.87), E20063 (0.87), E21447 (0.87), GERFT (0.85), E20066 (0.84), E20084 (0.84), GERFF (0.84).

Year	Top 10 variables (response rate in parenthesis)
1999	TVLSP (0.86), E20062 (0.84), E21423 (0.84), PFSI1 (0.84), E20063 (0.84), E21447 (0.84), GERFT (0.81), E20064 (0.81), E20066 (0.81), GERFF (0.81).

Table 18 shows the most frequent variables at the bottom 10 response rates from 1999 to 2009, and Table 19 shows the lists of the bottom 10 variables per year. Variables related to participation in tertiary education (student mobility), completion / graduation from primary education (graduation rates, total, male and female), teachers (percentage of trained teachers), and expenditure (public current expenditure on education as % of total current government expenditure) seem to be the least performing across years. The range of the bottom 10 response rates varies between 0.01 and 0.17.

Table 18. Variables at the bottom 10 response rates from 1999 to 2009 - case A

Concept	Subgroup	Parent	Code	Frequency
Students from a given country studying abroad (outbound mobile students)	Participation	Student mobility indicators	FSOABS	11
Outbound mobility ratio (%) – Tertiary education	Participation	Student mobility indicators	FSOPTE	11
Gross outbound enrolment ratio – Tertiary education	Participation	Student mobility indicators	FSOPTP	11
Expected gross primary graduation rate. Female	Completion	Completion / graduates ratios	EGGFF	10
Gender parity index for expected gross primary graduation rate	Completion	Completion / graduates ratios	EGGFG	10
Expected gross primary graduation rate. Male	Completion	Completion / graduates ratios	EGGFM	10
Expected gross primary graduation rate. Total	Completion	Completion / graduates ratios	EGGFT	9
Gender parity index for % of trained teachers. Upper secondary	Teacher	Trained teacher	GPTTU	8
Percentage of trained teachers. Upper secondary. Female	Teacher	Trained teacher	TRAUF	6
Percentage of trained teachers. Upper secondary. Male	Teacher	Trained teacher	TRAUM	6

Concept	Subgroup	Parent	Code	Frequency
Gross primary graduation rate. Female	Completion	Completion / graduates ratios	GGFF	5
Gender parity index for gross primary graduation rate	Completion	Completion / graduates ratios	GGFG	5
Gross primary graduation rate. Male	Completion	Completion / graduates ratios	GGFM	5
Public current expenditure on education as % of total current government expenditure	Expenditure	Public current expenditure	PCCGE	5
Gross primary graduation rate. Total	Completion	Completion / graduates ratios	GGFT	4
Gender parity index for % of trained teachers. Lower secondary	Teacher	Trained teacher	GPTTL	4

Table 19. List of bottom 10 variable response rates - case A

Year	Bottom 10 variables (response rate in parenthesis)
2009	EGGFF (0.03), EGGFG (0.03), EGGFM (0.03), EGGFT (0.03), GPSR4 (0.04), SR4FF (0.04), SR4FM (0.04), SR4FT (0.04), GPTR (0.04), TRANF (0.04).
2008	R25004 (0.14), EGGFF (0.15), EGGFG (0.15), EGGFM (0.15), EGGFT (0.16), R25000 (0.16), PCCGE (0.16), GPTTU (0.17), TRAUF (0.17), TRAUM (0.17).
2007	GPTTU (0.12), TRAUF (0.12), TRAUM (0.12), TRAUT (0.12), GPTTL (0.14), TRALM (0.14), PCCGE (0.15), TRALF (0.15), GPTTS (0.16), TRALT (0.16).
2006	EGGFF (0.13), EGGFG (0.13), EGGFM (0.13), PCCGE (0.13), GPTTU (0.13), TRAUF (0.13), TRAUM (0.13), TRAUT (0.14), EGGFT (0.15), GPTTL (0.15).
2005	EGGFF (0.09), EGGFG (0.09), EGGFM (0.09), EGGFT (0.1), GPTTU (0.13), TRAUF (0.13), TRAUM (0.13), GGFF (0.13), GGFG (0.13), GGFM (0.13).
2004	EGGFF (0.14), EGGFG (0.14), EGGFM (0.14), GPTT0 (0.14), GPTTU (0.15), TRAUF (0.15), TRAUM (0.15), EGGFT (0.16), PCCGE (0.16), GPTTL (0.16).
2003	PCCGE (0.11), GPTTL (0.11), GPTTU (0.11), TRALF (0.11), TRALM (0.11), TRAUF (0.11), TRAUM (0.11), EGGFF (0.12), EGGFG (0.12), EGGFM (0.12).
2002	EGGFF (0.06), EGGFG (0.06), EGGFM (0.06), EGGFT (0.07), GGFF (0.07), GGFG (0.07), GGFM (0.07), GGFT (0.08), GPTT0 (0.12), GPTTU (0.12).
2001	EGGFM (0.01), EGGFT (0.01), GGFF (0.01), GGFG (0.01), GGFM (0.01), GGFT (0.01), EGGFF (0.02), EGGFG (0.02), GPTT0 (0.12), GPTTU (0.13).
2000	EGGFM (0.01), EGGFT (0.01), GGFF (0.01), GGFG (0.01), GGFM (0.01), GGFT (0.01), EGGFF (0.01), EGGFG (0.01), XSINT (0.09), EC2TO (0.12).

1999	EGGFF (0.01), EGGFG (0.01), EGGFM (0.01), EGGFT (0.01), GGFF (0.01), GGFG (0.01), GGFM (0.01), GGFT (0.01), XSINT (0.04), XSPR0 (0.04).

3.4 Most frequent variables used in international reports

There are certain variables that are frequently part of education reports from international organizations. Annex 1 presents these variables, which correspond mainly to two sources: the World Bank education statistics report (World Bank, 2011a) and the UIS country profile report (UNESCO-UIS, 2011c). Additional variables that are of interest to education analysts (based on personal experience) are also presented. From these variables, 45 were selected as a representative sample of important variables used in education reports and analyses. Basically, this group comprises variables that represent a "total" (male and female) and if available, the gender parity index (GPI) of the respective variable. It is worth to mention that many variables from this group could be replaced by other variables - proxy variables - with equal importance related to interpretation or policy monitoring. The GPI represents the relationship between the values for the male and the female populations for a given indicator, and it can be obtained only if the disaggregated value by sex is reported.

Table 20 shows the comparison of the average response rate per year between case A and case B for these 45 selected variables. These averages are similar to those from the complete dataset (see Table 12). An important aspect to highlight is the effect of UIS estimations on the 45 selected variables versus the complete dataset. The average proportion from 1999 to 2009 of UIS estimations in the 45 selected variables is 7.95%, while the proportion of UIS estimations on the complete dataset is only 5.04%, possibly as a result of the greater attention that this group of variables receives in the UIS production of statistics.

Table 20. Basic statistics for the 45 selected variables

	Response rate filtered variables		Difference	Number of UIS estimations
	Submitted and UIS estimations (A)	Submitted data only (B)		
year	Mean	Mean	A-B	
1999	0.545	0.449	0.096	903
2000	0.556	0.443	0.113	1063
2001	0.572	0.458	0.114	1072
2002	0.590	0.485	0.105	988
2003	0.585	0.459	0.126	1185
2004	0.600	0.510	0.09	846
2005	0.588	0.508	0.08	752
2006	0.548	0.497	0.051	480
2007	0.582	0.534	0.048	451
2008	0.550	0.520	0.03	282
2009	0.415	0.394	0.021	198

CHAPTER 4. Statistical Capacity Indicator (SCI) and country response rates (CRR)

In this section we investigate the relationship between the process of capacity building for education statistics and the broader nature of countries' production of official statistics. It is relevant to mention that, in many if not the majority of cases, education statistics are handled by the component of the government involved directly in education (e.g. Ministry of education, higher education, etc.). Nevertheless, it is very possible that many country-level factors affect the national statistical production and the production of education statistics at the same time.

The Statistical Capacity Indicator (SCI) is an initiative from the Bulletin Board on Statistical Capacity (BBSC) and the Development Data Group (DECDG) at the World Bank that provides a country-level assessment, in the form of a composite score, of various important aspects of the capacity of national statistical systems from 145 developing countries. This multidimensional diagnostic framework is composed of 3 dimensions (World Bank, 2011b): statistical methodology, source data and periodicity and timeliness. A score of 0 indicates that the national statistical system does not meet any of the proposed criteria of the diagnostic framework, and 100 indicate that all criteria are fully satisfied; moreover, the average score for the 145 countries for 2004-2010 is 63 points. The source is: "Bulletin Board on Statistical Capacity, The World Bank" and the data are freely available from the World Bank website (BBSC 2011).

The SCI seeks to capture the different aspects of the national statistical systems, which the production of national education statistics could be considered part of. For this reason, the comparison of SCI scores and the country response rates (CRR) would give an important insight over the relationship between national statistical capacities and the completeness of UIS database.

Table 21 shows the correlation between the CRR (case A – submitted data and UIS estimations) and the scores from the SCI. The latter is only available for years 1999 and 2004-2010. It can be seen that the correlation and the variance explained are at their highest between 2004 and 2008, averaging 0.516 and 26.8% respectively. The consistent positive relationship between the aspects measured by the SCI score and the capacity of the country to respond to the UIS education questionnaire may indicate that the elements affecting national statistical systems, such as investment, political will and internal demand, may also have a significant effect on national education statistical systems.

Table 21. Correlation and variance explained – SCI scores and CRR (case A)

	1999	2004	2005	2006	2007	2008	2009
Correlation (r)	0.283	0.493	0.572	0.544	0.501	0.471	0.111
Variance explained (r^2) %	8.0	24.3	32.8	29.6	25.1	22.2	1.2

The fact that a country has relatively high (over 50 points) SCI score is significant in many ways. The SCI score is a multidimensional scale and the obtention of a good score in the ample range of the SCI evaluation criteria may imply the existence of a coordinated system that produces certain statistics under the demands or supervision of governmental officials. These countries suppose a good opportunity for building capacity for reporting education statistics, as some of the required conditions that are very difficult to set up and maintain may already be in place. Moreover, national officials may be more prone to investing in education statistics if they notice that their production is lagging behind their national statistical capability.

Figure 7 presents the plot of SCI scores and CRR for case A (submitted data and UIS estimations) for 145 countries for 2007 and 2008. It can be noted that most countries are located in the upper right quadrant (59% and 62% for 2007 and 2008, respectively), which indicates a moderate to good SCI score (over 50) and a moderate to good CCR (over 0.5). Nevertheless, there are 29 countries (20% of 145) in 2007 and 19 (13% of 145) in 2008 that fall within the lower right quadrant, which indicates a moderate to good statistical capacity (score over 50) but a low to moderate response to the UIS education survey (from 0 to 0.5).

For other years, the situation is almost the same: the proportion of countries in this quadrant (SCI > 50 and CCR < 50) for 2004 is 9%, for 2005, 14%, for 2006, 16% and for 2009, 30%. The analysis with CRR case B reveals a similar pattern, with an average of 7 more countries in the lower right quadrant each year.

Figure 7. SCI scores and CRR - case A

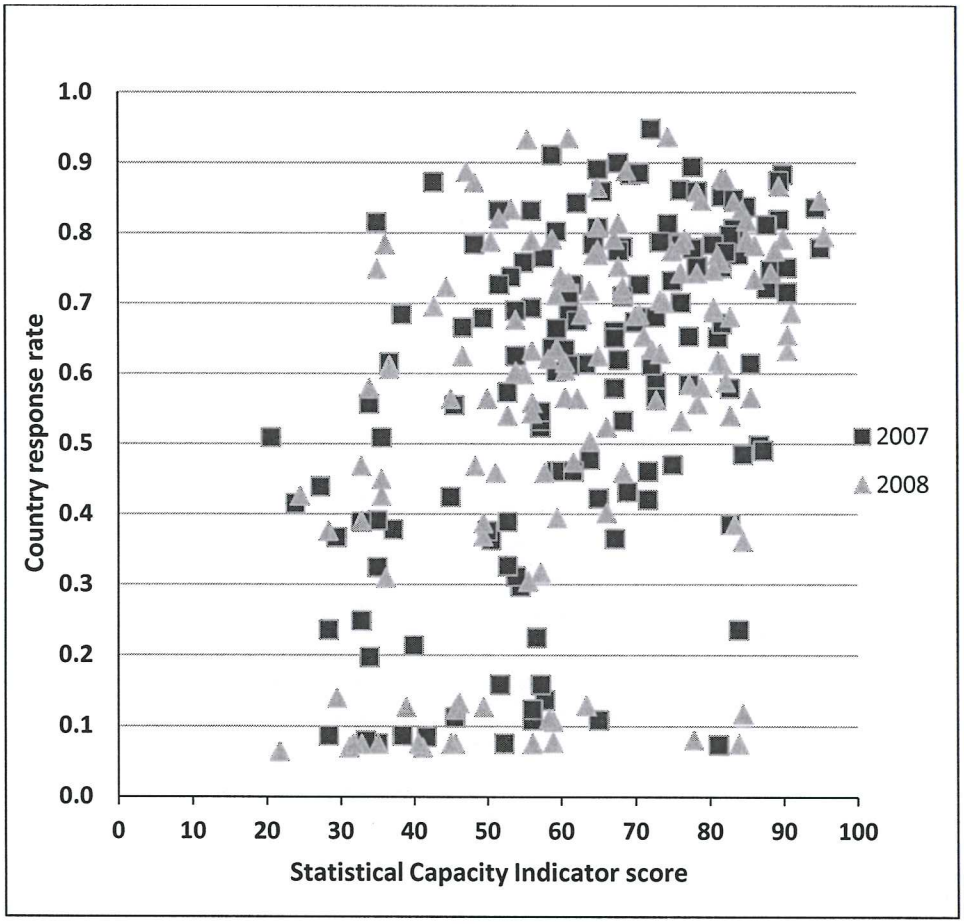


Table 22 presents the list of countries featuring high potential for statistical reporting, yet low response rate to the UIS education survey. Among the countries with average (from 2004 to 2008) high SCI score and average low CRR we found: Egypt (SCI 83 and CRR 0.415), Moldova (SCI 80 and CRR 0.157), Albania (SCI 78 and CRR 0.173), Slovak Republic (SCI 77 and CRR 0.259), etc.

Table 22. List of countries with average SCI score (>50) and low response rates (<0.5)

Country	Mean SCI 2004-08	Mean CRR 2004-08
Egypt, Arab Rep.	**83**	0.415
Moldova	**80**	0.157
Albania	**78**	0.173
Slovak Republic	**77**	0.259
Côte d'Ivoire	**73**	0.356
Nepal	**72**	0.398
Sri Lanka	**69**	0.454
Vietnam	**68**	0.366
Bolivia	**66**	0.499
Rwanda	**61**	0.423
China	**61**	0.430
Bhutan	**59**	0.499
Chad	**57**	0.488
St. Lucia	**57**	0.454
St. Vincent and the Grenadines	**57**	0.454
St. Kitts and Nevis	**57**	0.454
Zimbabwe	**56**	0.103
Yemen, Rep.	**55**	0.375
Montenegro	**55**	0.127
Benin	**54**	0.448
Tonga	**54**	0.318
Comoros	**53**	0.396
Bosnia and Herzegovina	**52**	0.168
Samoa	**51**	0.272

Figure 8 presents the histogram of the difference between country response rates (CRR) and SCI scores for 145 countries. For any given country, a positive difference indicates that the proportion of available responses in the UIS database is higher than the evaluation of the respective national statistical capacity (as measured by the CCR), while a negative difference indicates the contrary. It can be noted that most countries (84) are located in the negative side of the "CRR minus SCI" spectrum, with some countries of negative differences being at a relatively larger distance from the zero (compared to countries with positive differences). It can

be also noted that 50% of countries in the sample have their difference (CRR-SCI) in the range of -10 to 10.

Figure 8. Histogram of 100*(CRR minus SCI) / average 2004-08

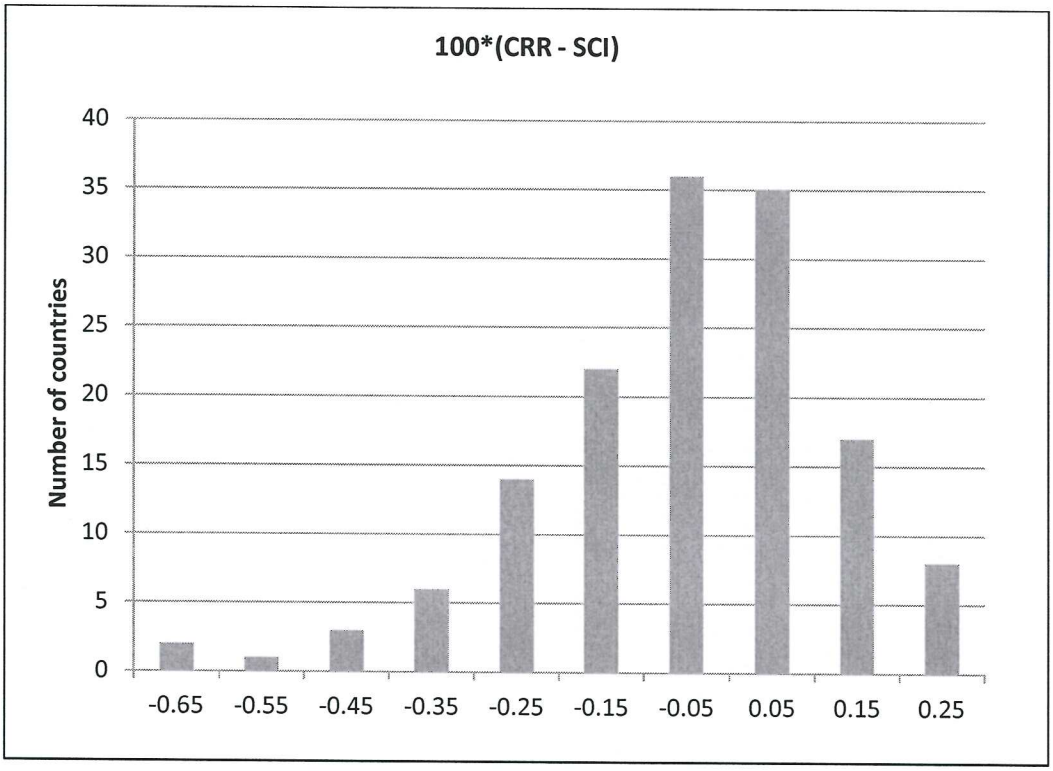


Table 23. List of countries with the largest negative difference between response rates and SCI scores

Country	100*(CRR – SCI) 2004-08
Turkmenistan	-34.9
Gabon	-30.9
Brazil	-29.6
Haiti	-29.3
Sierra Leone	-29.2
India	-28.7
South Africa	-28.1
Papua New Guinea	-27.0
Thailand	-25.3
Russian Federation	-25.3
Guinea-Bissau	-24.6
Jamaica	-21.9
Poland	-20.9
Angola	-19.3
Yemen, Rep.	-17.8

Note: excludes the countries already shown in Table 22

Table 23 shows a list of countries with the largest negative difference between CRR and SCI scores, in other words, countries that are doing worse than what it would be expected from their SCI scores (excludes the countries already shown in Table 22). As with countries in Table 22, these countries have also the potential for improving their report on education statistics.

Table 24. List of countries with the largest positive difference between response rates and SCI scores

Country	100*(CRR – SCI) 2004-08
Eritrea	44.8
Djibouti	28.3
Niger	26.5
Burundi	24.9
Lao PDR	22.6
Sudan	22.1
Belize	21.9
Dominica	21.1
Guinea	20.9
Madagascar	18.8
Sao Tome and Principe	17.5
Syrian Arab Republic	16.8
El Salvador	16.3
Mauritania	15.4
Cape Verde	15.2

Table 24 shows a list of countries with the largest positive difference between CRR and SCI scores, in other words, countries that are doing much better than what it would be expected from their SCI scores. Considering that most of these countries are developing economies, it is very possible that their relatively good performances in response rates are the result of political or institutional compromises toward the use of national education statistics.

In conclusion, the SCI assessment may be useful to point out countries where statistical capacity building of education statistics may be successful due to the presence of national statistical system of moderate or advance stage of efficiency. Moreover, the positive correlation between SCI scores and CRR may indicate common underlying factors affecting both, or cause and effect relationship, which could be exploited to increase UIS education response rates. Countries with relatively good response rates with respect to their SCI score are also worth further examination, as their statistical capacity for education statistics may be an example for other countries that face the same challenges (developing economies).

CHAPTER 5. Trajectory of response rates by subgroups and binary time series by variable

5.1 Response feature analysis by subgroups

For researchers, national officials, or international agencies who depend on the UIS education database, an important issue is to know if the completeness of the database is increasing (related to the increase of CRR or the decrease of missing values) or decreasing. In this regard, the following analysis of responses by subgroups of variables (as defined by UIS, based on the concept that the variables aim to measure) gives a richer look into the structure or patterns of the database missing values across time than the descriptive analysis of CRR averages.

There are 39 subgroups of variables as defined by UIS. A basic clustering of these subgroups can be made by carrying a linear regression of each subgroup response rate (dependent variable) on year (independent variable), and then dividing them into two segments: subgroups with negative slopes and subgroups with positive slopes. This approach is relatively straightforward and involves the use of regression coefficients as summary measures[1]. Because of the previously referred problems with the 2009 data collection, only data from 1999 to 2008 will be incorporated in the present analysis.

Figure 9 shows the subgroups that, based on the linear regression, exhibit negative slopes from 1999 to 2008, and Table 25 shows the average response rate of these subgroups. In absolute values, the maximum slope is 0.0115, which represent the biggest decrease (related to "Distribution of tertiary students") and the minimum is 0.0031, which is very close to zero (related to "Survival"). The total number of variables in this category (subgroups with negative slopes) is 161, representing 30% of the variables in the analysis.

In order to assess the impact that a decreasing trend may have, information about the suggested lost of data points per year is presented (see Table 25), which is the result of the slope multiplied by the quantity of variables in the subgroup. We can note that the subgroups that may be more sensible to data loss are: Teaching staff (74 data points lost per year), Gross enrolment ratio (40), Repetition rates (31), Enrolment in tertiary (28) and School life expectancy (25). Under the linear regression model, the total number of lost data point per year is 258.

---

[1] For other approaches related to *response feature analysis*, see Der and Everit (2002).

Figure 9. Response rates by variable subgroups (case A) - negative slope

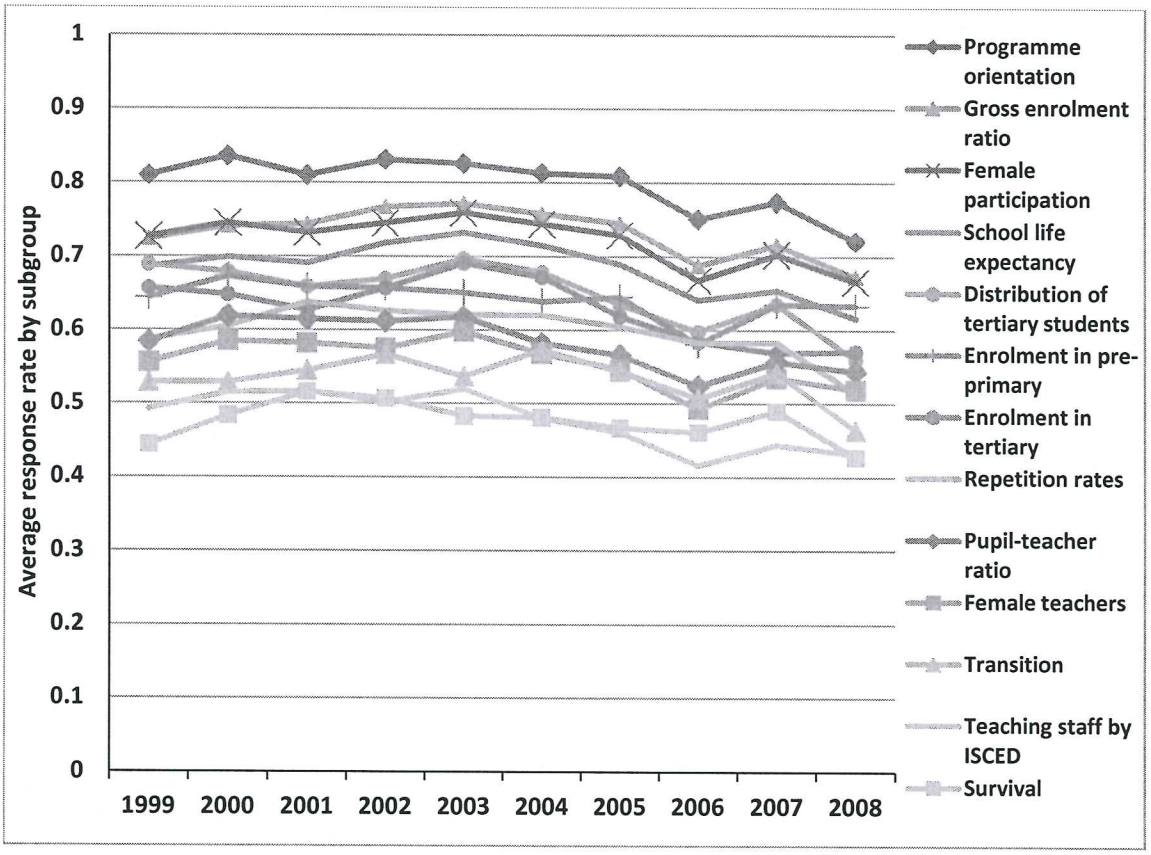


Table 25. Average response rate by subgroup from 1999 to 2008 – negative slopes

Subgroup (number of variables in parenthesis)	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	Mean 99-08	Lost data points per year
Programme orientation (3)	0.81	0.84	0.81	0.83	0.83	0.81	0.81	0.75	0.77	0.72	0.80	6.11
Gross enrolment ratio (32)	0.72	0.74	0.74	0.77	0.77	0.76	0.75	0.69	0.72	0.67	0.73	**40.29**
Female participation (10)	0.73	0.75	0.73	0.75	0.76	0.74	0.73	0.67	0.70	0.67	0.72	15.53
School life expectancy (15)	0.69	0.70	0.69	0.72	0.73	0.72	0.69	0.64	0.65	0.62	0.68	**24.62**
Distribution of tertiary students (3)	0.69	0.68	0.66	0.67	0.70	0.68	0.64	0.60	0.64	0.56	0.65	7.21
Enrolment in pre-primary (4)	0.64	0.67	0.66	0.66	0.65	0.64	0.65	0.58	0.64	0.63	0.64	4.07
Enrolment in tertiary (15)	0.66	0.65	0.63	0.66	0.69	0.67	0.62	0.58	0.57	0.57	0.63	**27.58**
Repetition rates (21)	0.59	0.61	0.64	0.63	0.62	0.62	0.61	0.58	0.58	0.51	0.60	**30.59**
Pupil-teacher ratio (5)	0.58	0.62	0.62	0.61	0.62	0.58	0.57	0.52	0.56	0.54	0.58	9.01

Subgroup (number of variables in parenthesis)	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	Mean 99-08	Lost data points per year
Female teachers (5)	0.56	0.59	0.58	0.58	0.60	0.57	0.55	0.49	0.54	0.52	0.56	7.87
Transition (4)	0.53	0.53	0.55	0.57	0.54	0.57	0.54	0.51	0.54	0.46	0.53	3.63
Teaching staff by ISCED (34)	0.49	0.52	0.52	0.50	0.52	0.48	0.46	0.42	0.44	0.43	0.48	**74.20**
Survival (12)	0.44	0.48	0.52	0.51	0.48	0.48	0.47	0.46	0.49	0.43	0.48	7.72

Figure 10 shows the subgroups that exhibit positive slopes from 1999 to 2008, and Table 26 shows the average response rates of these subgroups. The steepest slope (the most increasing trend) is 0.019, related to "Repeaters in secondary" and the minimum is 0.000073, which are very close to zero (related to "School age"). The total number of variables in this category (subgroups with positive slopes) is 381, representing 70% of the variables in the analysis.

We can note that the subgroups that seem to be gaining more data points per year are: Percentage of repeaters (101 data points gained per year), Repeaters in secondary (89), Enrolment in secondary by grade (87), Completion / graduates ratios (44), Percentage of GDP / GNP (29), New entrants (20), Educational expenditure by nature (21) and Public current expenditure (16). Under the linear regression model, the total number of gained data point per year is 524 and the difference between the gained and lost data points per year is 265, in other words, an addition of 265 data points is expected each year. Taking into account that the total number of data points on consideration is 542x209 = 113278, the net gain of data points would represent only the 0.23% of this dataset.

Figure 10. Response rates by variable subgroups (case A) - positive slope

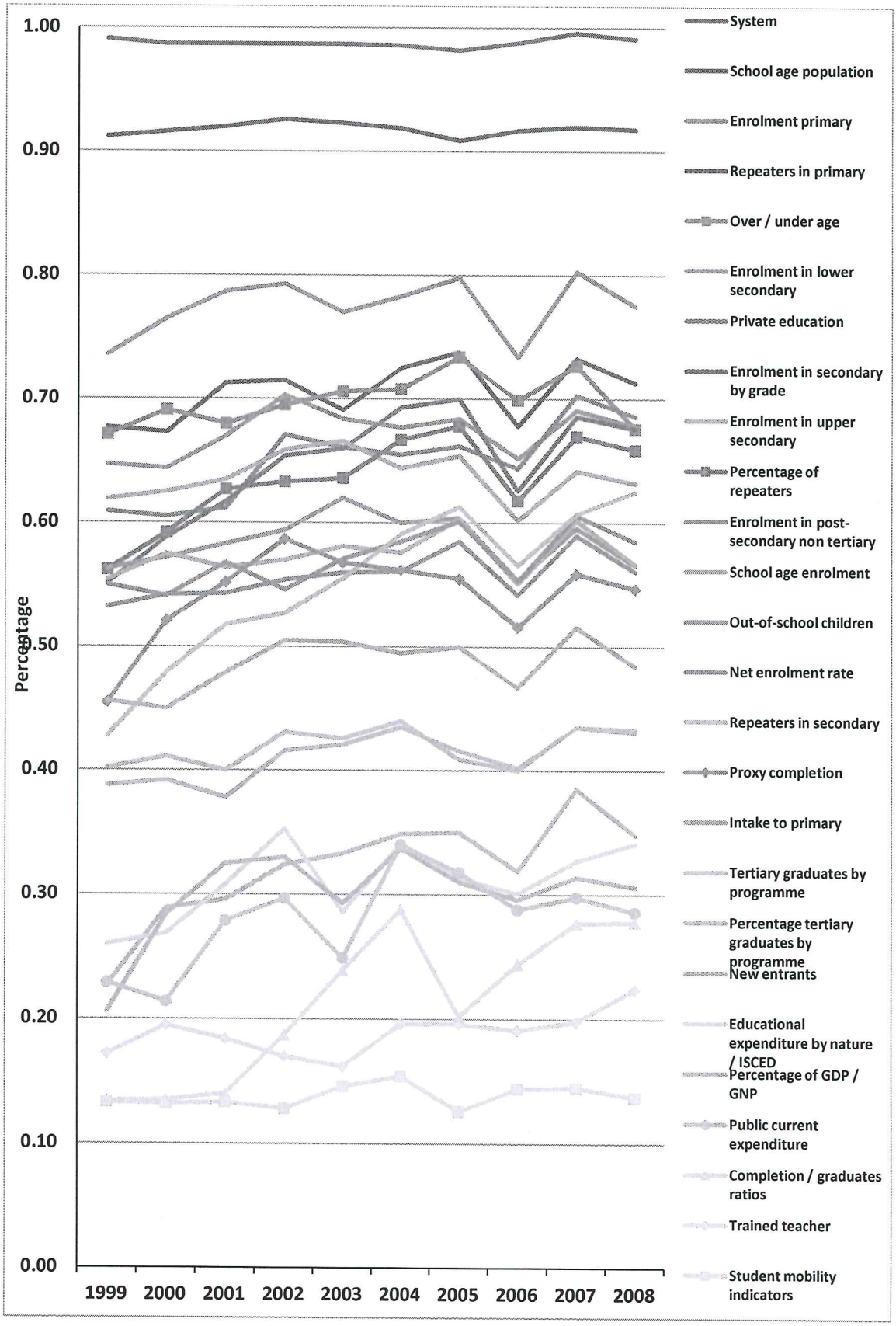


Table 26. Average response rates by subgroup from 1999 to 2008 – positive slopes

Subgroup (number of variables in parenthesis)	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	Mean 99-08	Gained data points per year
System (17)	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99	1.00	0.99	0.99	1.12
School age population (24)	0.91	0.92	0.92	0.93	0.92	0.92	0.91	0.92	0.92	0.92	0.92	0.36
Enrolment primary (22)	0.74	0.77	0.79	0.79	0.77	0.78	0.80	0.73	0.80	0.78	0.77	10.59
Repeaters in primary (18)	0.68	0.67	0.71	0.72	0.69	0.73	0.74	0.68	0.73	0.71	0.71	15.32
Over / under age (6)	0.67	0.69	0.68	0.70	0.71	0.71	0.73	0.70	0.73	0.68	0.70	3.88
Enrolment in lower secondary (12)	0.65	0.64	0.67	0.70	0.68	0.68	0.68	0.65	0.69	0.68	0.67	6.90
Private education (7)	0.61	0.61	0.61	0.67	0.66	0.66	0.66	0.64	0.70	0.69	0.65	13.35
Enrolment in secondary by grade (34)	0.55	0.59	0.62	0.65	0.66	0.69	0.70	0.63	0.69	0.68	0.65	87.30
Enrolment in upper secondary (12)	0.62	0.63	0.64	0.66	0.67	0.64	0.65	0.60	0.64	0.63	0.64	0.52
Percentage of repeaters (52)	0.56	0.59	0.63	0.63	0.64	0.67	0.68	0.62	0.67	0.66	0.63	101.43
Enrolment in post-secondary non tertiary (4)	0.56	0.57	0.58	0.59	0.62	0.60	0.60	0.55	0.60	0.57	0.59	0.30
School age enrolment (9)	0.55	0.58	0.56	0.57	0.58	0.58	0.60	0.55	0.60	0.57	0.57	3.82
Out-of-school children (6)	0.55	0.54	0.57	0.55	0.57	0.59	0.60	0.55	0.61	0.59	0.57	6.64
Net enrolment rate (16)	0.53	0.54	0.54	0.55	0.56	0.56	0.59	0.54	0.59	0.56	0.56	13.90
Repeaters in secondary (22)	0.43	0.48	0.52	0.53	0.56	0.59	0.61	0.57	0.61	0.63	0.55	89.26
Proxy completion (8)	0.46	0.52	0.55	0.59	0.57	0.56	0.56	0.52	0.56	0.55	0.54	8.23
Intake to primary (14)	0.46	0.45	0.48	0.51	0.50	0.50	0.50	0.47	0.52	0.48	0.49	11.17
Tertiary graduates by programme (2)	0.40	0.41	0.40	0.43	0.43	0.44	0.41	0.40	0.44	0.43	0.42	1.00
Percentage tertiary graduates by programme (1)	0.39	0.39	0.38	0.42	0.42	0.44	0.42	0.40	0.44	0.43	0.41	1.04
New entrants (8)	0.23	0.29	0.30	0.32	0.33	0.35	0.35	0.32	0.39	0.35	0.32	19.96
Educational expenditure by nature / ISCED (16)	0.26	0.27	0.31	0.35	0.29	0.34	0.31	0.30	0.33	0.34	0.31	20.98
Percentage of GDP / GNP (24)	0.21	0.28	0.33	0.33	0.29	0.34	0.31	0.30	0.31	0.31	0.30	28.85

Subgroup (number of variables in parenthesis)	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	Mean 99-08	Gained data points per year
Public current expenditure (10)	0.23	0.21	0.28	0.30	0.25	0.34	0.32	0.29	0.30	0.29	0.28	16.47
Completion / graduates ratios (12)	0.13	0.14	0.14	0.19	0.24	0.29	0.20	0.24	0.28	0.28	0.21	44.14
Trained teacher (20)	0.17	0.20	0.18	0.17	0.16	0.20	0.20	0.19	0.20	0.22	0.19	16.11
Student mobility indicators (5)	0.13	0.13	0.13	0.13	0.15	0.15	0.13	0.14	0.15	0.14	0.14	1.17

One critical assumption of the linear regression model is that the disturbance terms have zero covariance (Johnston, J., 1960), which for time series data equals to serial independence of the error (disturbance) terms. The present case of national statistical capacity and yearly responses to education surveys seems to be a circumstance where autocorrelated disturbances are very plausible. In this case, the estimations of the beta coefficients (slopes) remain unbiased, but the estimations of the sampling variances of these coefficients are likely to be seriously underestimated and, due to this, no longer valid. Therefore, it is more difficult to assess which subgroups have slopes that are significantly different from zero.

5.2 Control charts of response rate by subgroups

Instead of assuming linear trends and relying on the significance test for the slope, we look at a different approach to evaluate the stability of the response rates.

The idea of assessing the statistical capacity through a score (SCI) was presented in Chapter 4. From this perspective, the production of international education statistics can be understood as a process with the capability of elaborate education statistics (output) that meets diverse user's needs (e.g. education planning departments, policy makers, researchers, general public, etc.). Borrowing from the theory of statistical quality control, capability analysis (control charts) could be used to monitor the process variability related to data production (presence or absence of data in the database) and to assess if this process is in statistical control. Statistical control, as defined by Montgomery (2001), is when the causes of variation of a process are due only to chance, in other words, only the natural variability that is caused by the cumulative effect of small unavoidable causes is present. In the case of statistics production, response rates in statistical control can vary around the mean due to natural variability (e.g. human errors in filling the questionnaire, etc.). Nevertheless, other causes of variability can be present in the output. Sometimes these are usually large compared to chance causes, and could be considered

as unacceptable level of performance by the part of the production process (or as remarkable national or international efforts). They are normally called "assignable causes" and, when present, the process is considered to be operating "out of control". In the case of statistical production, among the positive assignable causes (causes that make the response rate to increase beyond natural variability limits), we could find: heavy investment on education statistics by national or international sources (e.g. training, resources, etc.), UIS missions on capacity building; and among the negative assignable causes we could find: national authorities have decided to stop reporting data due to political issues, data on population is not suitable for the production of reliable statistics, etc.

To monitor subgroup's response rate performance and to differentiate between a process of education statistics production in statistical control and a process out of control and are important activities for data collection authorities because they help taking corrective actions and proposing adequate strategies looking for to improve the completeness of the education database. Control charts for the mean and the variance are well-known tools that allow comparing two output characteristics - in this case the mean response rate and the variance of the response rates - in each point of time to the expected natural variance of the process, represented by limits around the mean.

It is also important to note that the standard assumptions for the use of control charts are that the generated data, under statistical control, are normally and independently distributed. If we consider that each variable in analysis is the average of 209 observations and that sample sizes greater than 10 most of the times, then the normality condition may not be a problem (Central limit theorem). However, the assumption of independent observations (response rates) across years does not seem to be satisfactory. As mentioned in the case of linear regression, it is very plausible that response rates are correlated. This implies that, although the process can still be in control, the process mean is not invariable – it is continuously wandering. It is worth mention that positive autocorrelation will decrease the width of the control limits, and negative autocorrelation will increase them (Thaga, K., 2008). Positive autocorrelation means that a relatively large value in an observation is followed by another observation with a relatively large value, and a relatively small value is followed by another relatively small value, and negative autocorrelation means that one relatively low value is followed by a relatively high value, and vice versa. Based on this, the statistical production process may be hypothesized as displaying positive autocorrelations. Indeed, it is more possible that large values in response rates are followed by other large value, simply because there is a curve of learning which affect the

process positively. By the same token, a small value in response rates may be followed by another small value, as the system may be deteriorating due to lack of investment, reduction of personnel, etc. Because positive autocorrelation makes the limits narrower, the control charts will have the tendency to signal more data points out of control than there may be. Therefore, we could see the chart as a conservative tool to control processes.

From an exploratory point of view, control charts for the mean and the variance were applied to the response rate by subgroups, setting limits at three standard deviations (3-σ). Responses were previously defined as binary data (available or missing), probably coming from a binomial distribution; nevertheless, each variable in the analysis was the average of the response for the 209 countries, which suppose a good approximation to the normal distribution. As seen in Table 25 and 26, the sample size are constant within subgroups, but different subgroups may have different sample sizes (number of variables within each subgroup). The number of samples used in the construction of the limits that describes a process in statistical control is 10 (from 1999 to 2008). It is usually recommended to have at least 25 samples in order to calculate the limits, but in reality these limits can be calculated with small amount of data, the only problem being the reliability of the calculation of limits. In addition, it would take 15 more years to obtain 25 samples.

Based on the visual examination of the control charts for 38 subgroups (the subgroup "percentage in tertiary" has only 1 element and was excluded of the present analysis), three situations could be described: response rate production possibly in statistical control, possibly decreasing and possibly increasing (see Table 27). It can be noted that the majority of subgroups (25) fall in the category of "in control", while five subgroups seem to have an increasing trend and eight a decreasing trend.

Table 27. Results of the examination of control charts by subgroups

Subgroup	Number of variables	Description
Female participation	10	In control
Female teachers	5	In control
Intake to primary	14	In control
Net enrolment rate	16	In control
New entrants	8	In control
Out-of-school children	6	In control
Percentage of GDP / GNP	24	In control
Programme orientation	3	In control
Proxy completion	8	In control

Subgroup	Number of variables	Description
Public current expenditure	10	In control
Pupil-teacher ratio	5	In control
School age enrolment	9	In control
School age population	24	In control
Student mobility indicators	5	In control
System	17	In control
Teaching staff by ISCED	34	In control
Tertiary graduates by programme	2	In control
Trained teacher	20	In control
Completion / graduates ratios	12	In control
Distribution of tertiary students	3	In control
Educational expenditure by nature / ISCED	16	In control
Enrolment in lower secondary	12	In control
Enrolment in post-secondary non tertiary	4	In control
Enrolment in pre-primary	4	In control
Enrolment in secondary by grade	34	In control
Enrolment in tertiary	13	Decreasing
Enrolment primary	22	Decreasing
Gross enrolment ratio	32	Decreasing
Repetition rates	21	Decreasing
School life expectancy	15	Decreasing
Survival	12	Decreasing
Teaching staff by ISCED	34	Decreasing
Transition	4	Decreasing
Percentage of repeaters	52	Increasing
Private education	7	Increasing
Repeaters in primary	18	Increasing
Repeaters in secondary	22	Increasing
Over / under age	6	Increasing

For subgroups described as possibly in control, there was no evidence, based on their behaviours in the control charts, that their mean and variance were affected by assignable causes of variability. Figure 11 and 12 shows two examples of subgroups possibly in control (School age population and Distribution of tertiary students). It can be noted that the production of statistics related to School age population is very stable. The production of statistics related to Distribution of tertiary students seems to one or two point in a downward trend; nevertheless, there is no evidence that this production process is displaying behaviours out of control or not expected when natural variability is taken into account.

Figure 11. Control chart for response rate of "School age population"

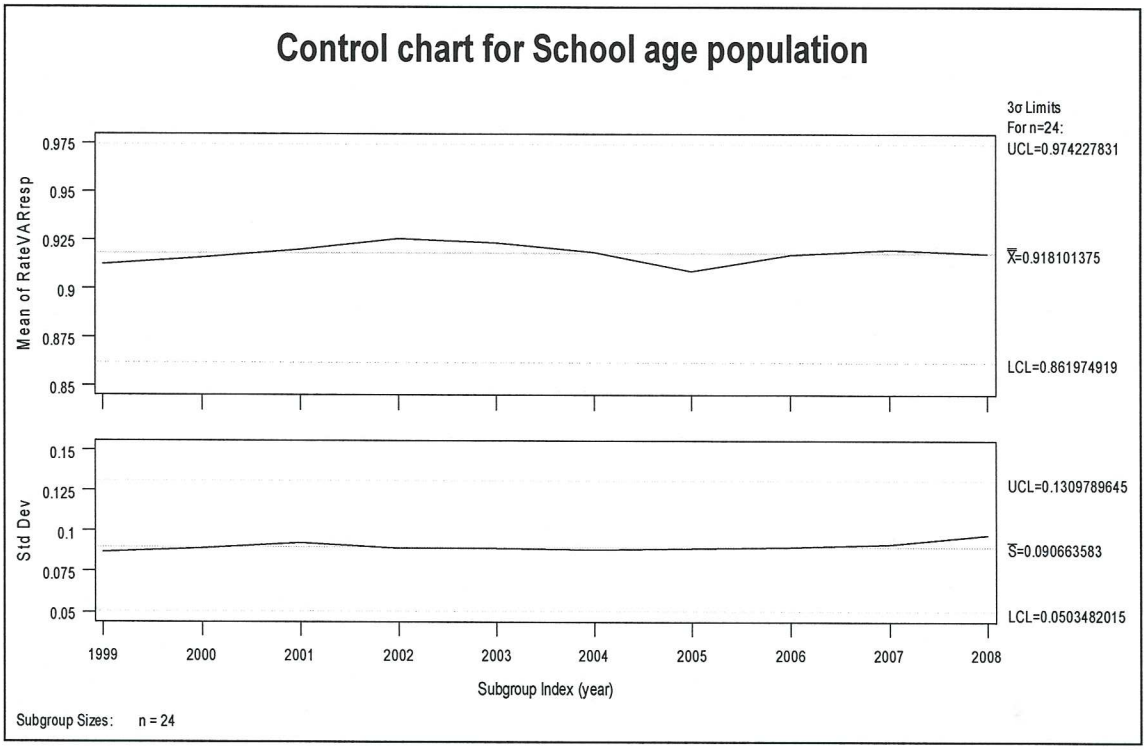


Figure 12. Control chart for response rate of "Distribution of tertiary students"

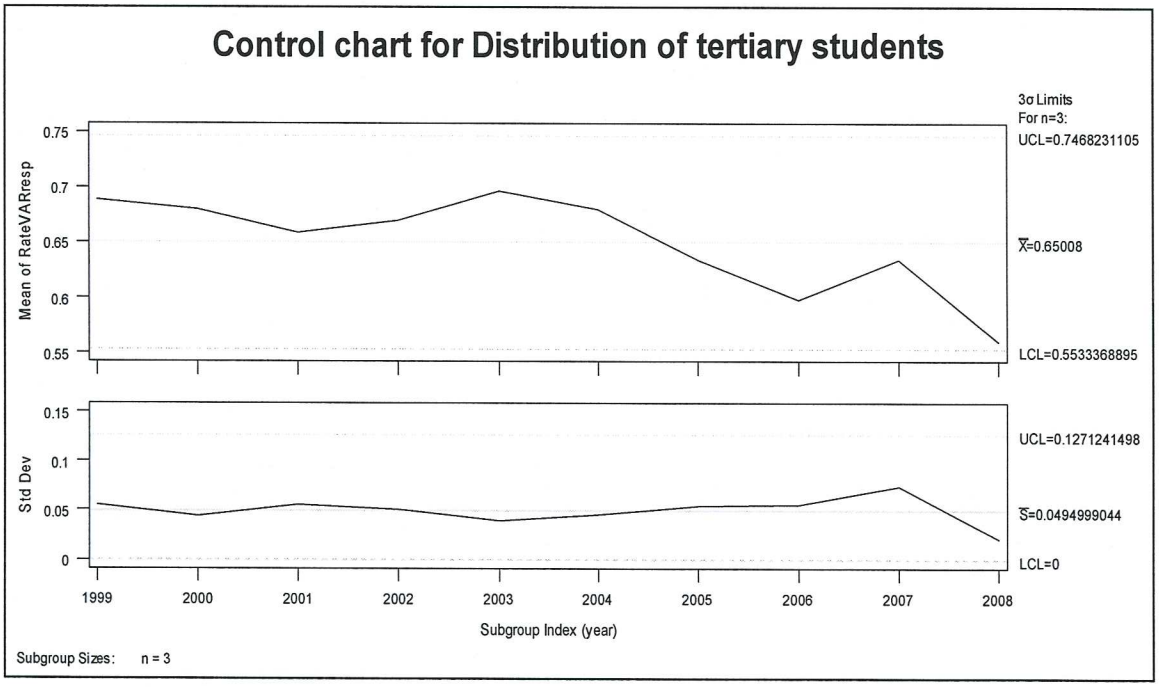


Figure 13 shows the control chart for Repeater in secondary. In the mean chart, we can notice an increasing trend of the average response rate since 1999 (below the lower control limit of 0.515)

to 2008 (over the upper control limit of 0.587), which could not be explained only by the natural variability of the production of this type of statistics.

Figure 13. Control chart for response rate of "Repeaters in secondary"

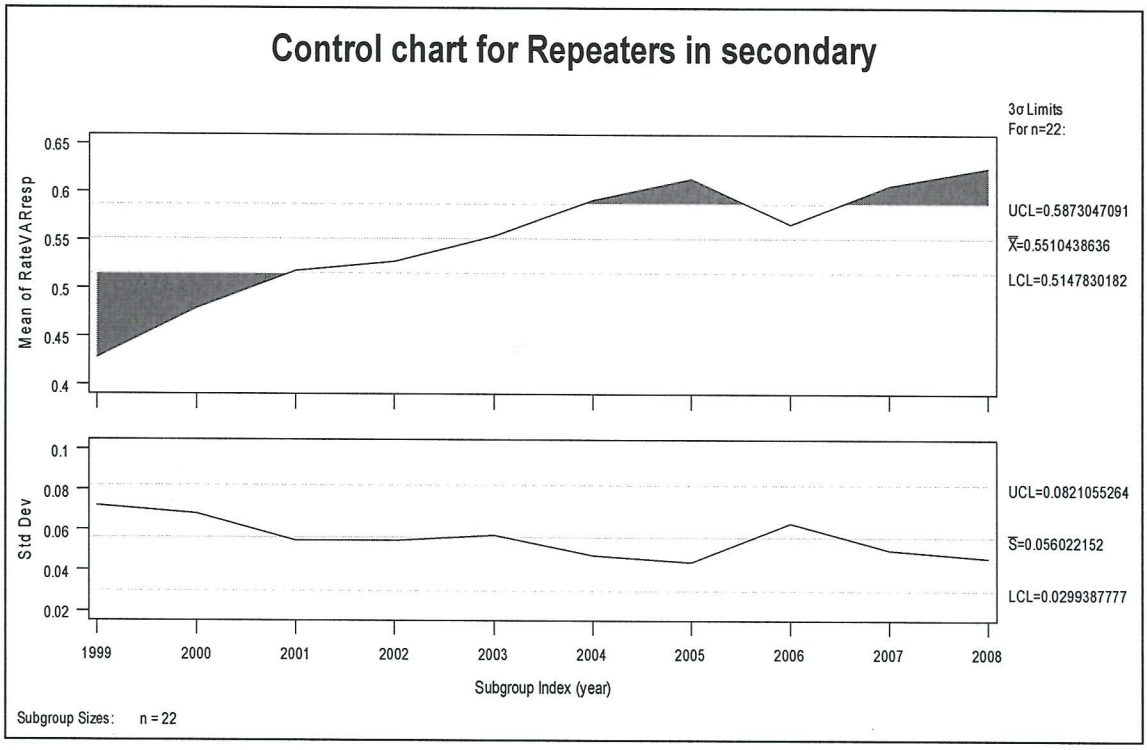
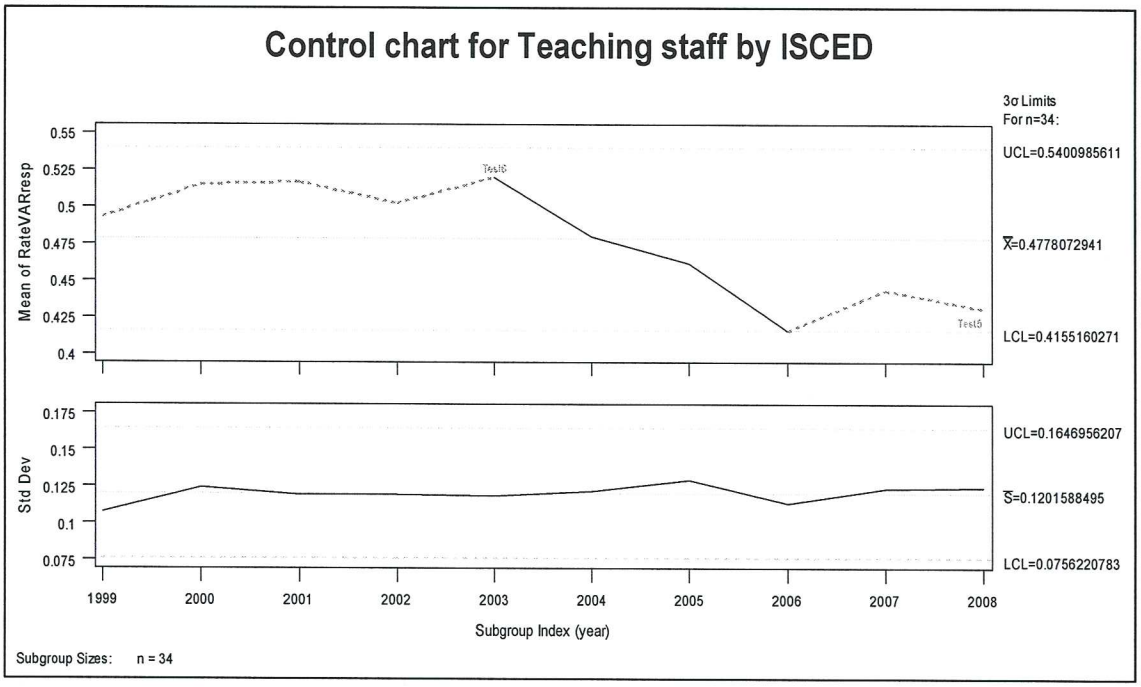


Figure 14 shows the control chart for Teaching staff by ISCED. In the mean chart, we can notice a decreasing trend of the average response rate between 2003 and 2006. Before and after this period, the response rate to this subgroup seems to be stable.

Figure 14. Control chart for response rate of "Teaching staff by ISCED"



Test 5: two to three points in a row in Zone A and beyond (Zone A: two to three standard deviations from the mean)
Test 6: four to five points in a row in Zone B or beyond (Zone B: one to two standard deviations from the mean)

In both cases (Repeaters in secondary and Teaching staff by ISCED), the causes that may be affecting the production of statistics are not revealed by the control charts; however analysts can use these tools to start an investigation and point out which actions are having a positive or negative impact in the production of statistics. Similarly, investigations could be undertaken to identify the causes of improvements for variables showing a significant increase.

As shown in the present analysis, control charts can be used to detect variations and trends in the production of responses while taking into account the natural variability of the process. For the case of production of statistics by subgroups, it can be concluded that the majority of them are stable processes (in control). If the aim is to increase the response rates (or decrease the number of missing values), it is necessary to design strategies that make subgroups "in control" display the positive trends of "increasing" response subgroup.

The detection of points (response rates) over or below the control limits, as well as non random patterns within the control limits, can guide future research aiming to understand which actions have positive or negative effects on the production of education statistics.

5.3 Analysis of binary time series

Section 3.1.5 describes a matrix of response as the matrix - consisting of zeroes and ones – that describes the absence or presence of observations on the education database (zero if the value is absent, one if it is present). We can apply the same transformation for the observations of a variable across time in order to obtain a vector that correspond to a binary time series, with each element representing a collection year. A simple analysis of these binary time series consists in the study of one-step transition counts (Cox and Snell, 1989). There are four transition counts of first order: 1) the number of times in a binary series that an element of value zero is followed by an element of value one (r_{01}), 2) an element of value one followed by zero (r_{10}), 3) element of value zero followed by zero (r_{00}), and 4) an element of value one followed by one (r_{11}). The last two cases (r_{00}, r_{11}) represent the circumstances where an observation is absent or present, respectively, during two consecutive years. The first case (r_{01}) corresponds to the circumstance where an observation is missing (absent) a given year, but the following year the observation is reported (present). The second case (r_{10}) is the most interesting: it represents the circumstance where an observation is present a given year, but it is not reported the following year. In other words, it denotes possible inconsistencies or difficulties in the continued production of a variable or statistic. A simple method to generalize the analysis of the second case in the education database is to sum the counts of each country time series by variable, and then to transform the total count into transition proportions. The transition proportion for the second case can be defined as:

$$total\ p_{10}\ of\ a\ variable = (sum\ of\ r_{10}\ from\ all\ countries) / (sum\ of\ r_{10}\ from\ all\ countries + sum\ of\ r_{11}\ from\ all\ countries).$$

Table 28 shows a list of the 20 variables with the highest transition proportion related to p_{10}. It can be noted that the most affected variables by inconsistencies in reporting across time are related to primary graduation (3 variables) and expenditure in education (17 variables). In general, these variables show a high count r_{00} (in average 1563, while the maximum r_{00} is 2090[2]), which is related to the inability to produce a valid value in two consecutive years. The average transition proportion p_{10} of all variables is 0.158, while the average p_{10} of the 20 variables in Table 29 is 0.358.

[2] 2090 = 209 countries multiplied 11 years minus 209. The subtraction of 209 corresponds to the ending effect of transition counts.

Table 28. List of the 20 variables with the highest transition proportion P_{10}

Name	R_{00}	R_{01}	R_{11}	R_{10}	P_{10}
Expected gross primary graduation rate. Total	1795	91	117	87	0.426
Expected gross primary graduation rate. Female	1820	82	110	78	0.415
Gender parity index for expected gross primary graduation rate	1820	82	110	78	0.415
Expected gross primary graduation rate. Male	1823	81	109	77	0.414
Educational expenditure by nature of spending as a % of total educational expenditure on public institutions. ISCED 1,2,3,4. Other current expenditure	1466	157	294	173	0.370
Educational expenditure by nature of spending as a % of total educational expenditure on public institutions. ISCED 1,2,3,4. Capital	1420	173	313	184	0.370
Educational expenditure by nature of spending as a % of total educational expenditure on public institutions. ISCED 1,2,3,4. Salaries	1470	154	296	170	0.365
Total expenditure on educational institutions and administration as a % of GDP. International sources. All levels	1507	164	267	152	0.363
Educational expenditure by nature of spending as a % of total educational expenditure on public institutions. ISCED 1,2,3,4. Total current expenditure	1426	167	318	179	0.360
Public current expenditure on education as % of total current government expenditure	1673	108	198	111	0.359
Total expenditure on educational institutions and administration as a % of GDP. All sources. Tertiary	1559	141	251	139	0.356
Total expenditure on educational institutions and administration as a % of GDP. All sources. All levels	1563	135	257	135	0.344
Total expenditure on educational institutions and administration as a % of GDP. All sources. Secondary and post-secondary non-tertiary	1579	132	249	130	0.343
Total expenditure on educational institutions and administration as a % of GDP. All sources. Pre-primary	1573	131	258	128	0.332
Percentage distribution of public current expenditure on education by level. Pre-primary	1268	193	423	206	0.328
Percentage distribution of public current expenditure on education not allocated by level	1154	220	483	233	0.325
Total expenditure on educational institutions and administration as a % of GDP. Private sources. Secondary and post-secondary non-tertiary	1587	125	256	122	0.323
Total expenditure on educational institutions and administration as a % of GDP. Private sources. Pre-primary	1592	124	254	120	0.321
Total expenditure on educational institutions and administration as a % of GDP. Private sources. Tertiary	1568	131	266	125	0.320
Total expenditure on educational institutions and administration as a % of GDP. All sources. Primary	1587	124	258	121	0.319

Based on the results of the analysis of time series, we could conclude that, these statistics - primary graduation and expenditure on education - may be suffering from a problem related to the consistency or reliability of reporting in addition to having relatively low response rates (expenditure by nature and completion/graduation ratios, respectively; see Table 26). The statistics display a particular challenge: the underlying capacity for production may exist, but the continuity of production is not assured. The causes of problems of reliability in the production/report can be multiple: lack of compromise from national authorities, statistics too costly to produce, lack of expertise on how to use them, etc. There is also the possibility that the distinct nature of these problematic statistics can be affecting the reliability of their production: primary graduates statistics are usually collected after the academic year is closed, and data on education expenditure usually come from the budget planning office, different in function and expertise than the office in charge of primary/secondary data collection.

CHAPTER 6. Underlying structure of responses – Factor analysis

The analysis or description of response rates (or missing values) for the average of 542 variables can overlook the response patterns of relevant group of variables while the analysis of the 39 subgroups can be a considerable task to handle; as a consequence, a better understanding of the patterns of response rates is needed.

This section presents the results of a factorial analysis that allows the identification and interpretation of dimensions related to the response matrix and the construction of scores on the proposed underlying dimensions. This will allow the reduction of the data to be examined and a better description and understanding of the behaviour of the response rates. From an exploratory perspective, factor analysis techniques will be use to identify or suggest an underlying structure of the response matrix.

6.1 Objective

The primary objective is to identify the structure in the response rates matrices and to compare them across years. This will allow to group variables and to reduce their number to more parsimonious set of data. The assumption is that certain underlying constructs exist, which affect the production of international comparable education statistics. The suggestion of a composite measure that summarizes the proposed factors will also prove helpful in posterior analyses.

6.2 Statistical model

The response matrix, as defined before, is a set of binary data, but factor analysis is designed for quantitative variables (Larocque, 2006). To circumvent this issue, the factor analyses were applied using tetrachoric correlations. The tetrachoric correlation supposes that any "two binary variables come from an underlying bi-normal model" and, in this regard, "the coefficient of tetrachoric correlation is an estimation of the correlation coefficient of the underlying bi-normal coefficient" [translated from Larocque (2006) : 76]. Regarding binary data describing any of the two states value/missing-value in the education dataset, the assumption that variables follow an underlying normal distribution may imply that there is an information threshold[3] in the process of data production by the side of the country respondents, and that a value is reported just when the information produced is above the threshold. This would be the case if, in order to report a data, country respondents must work estimations based on national data. Once they have all the raw material to produce a data point, it will be reported in the education survey. Nevertheless,

[3] In reality, the threshold will include factors like the quantity of hours-person available for the task or any other resource needed to complete the education survey.

the assumption of normal distribution of the variables could not be correct or hold for all situations, for example, if the country decides not to report any data when a minimum number of data points could not be produced, or if the country officials simply decided not to report any data, without consideration of the information that exist at the national level. However, with all these possible setbacks, the results obtained through the use of the tetrachoric correlation revealed an interesting data structure, which could indicate that, despite the inconveniences in national reporting cited before, the underlying assumptions may hold in this particular case.

To obtain the tetrachoric correlation matrix, the macro %POLYCHOR (SAS Institute Inc., 2005) from SAS was used (it took around 2-3 hours to complete the calculation for each response matrix). The parameters were: convergence = 0.00000001, maxiter = 50 and type = CORR. The default convergence parameter (0.0001) produced many missing values when processing the response matrix, especially in correlation values in the output matrix closer to 1, which is why it is recommended to use the parameter suggested above.

Regarding the size of the sample, there are several rules of thumb. Larocque (2006) recommends that the number of observations must be at least 5 times the number of the variables (ratio subject to item >= 5:1). In the case of analysing the entire response matrix, the ratio subject to item is 1: 2.5, well below the previous recommendation. Hair *et al.* (2009) also mention that, preferably, the number of observations must be larger than 100. The response matrix has 209 observations each year.

MacCallum *et al.* (1999) emphasize that the minimum number of observations or the minimum ratio subject to item depends on aspect of the inhered structure of the data and the design of the study. They consider that one of the main issues is related to the sampling errors and the recovery of population factors. Due to this, important aspects to take into consideration in order to determine the number of observations needed for a good factor extraction are high communalities (greater than 0.6), well-defined factors, overdetermination (small number of factors with many indicators each) and sample size. They remark that sample size will play an important role in factor recovery when the others aspects are not performing well.

Taking into account the previous discussion, the following strategy was implemented: application of factor analysis to the entire response matrix and to the 45-variable sample (see section 7.4) and then the comparison of results regarding the retained structure, the overall fit and the behaviour of the variables between these two input data (entire matrix versus 45 selected items). A scale for interpreting the retained dimensions is suggested based on the 45 previously selected variables, but this scale was improved through the deletion of some variables that were

loading in more than one dimension, and through the addition of other variables with high loading only in a given dimension. The results were tested in the response matrices of recent years in order to assess the stability of the structure retained.

As mentioned before, response matrices containing data for 522[4] variables (entire set) and a selected set of variables (totalling 45) with 209 observations are analyzed. However, variables with responses rate close to 1, such as age of entrance or duration of programmes as well as variables with response rate of zero, such as gross outbound enrolment ratio, outbound mobility ratio and outbound mobile students, were not included in the analysis (in total, 23).

6.3 Factor extraction and assessment of the models: results and comments

The software used in these analyses was SAS. As recommended by Larocque (2006), factor analysis with the method "common factor analysis" was applied. In SAS, this method is also known as "iterated principal factor analysis – PRINIT". Due to the presence of communalities higher than 1, in both the matrix with 522 variables and in the selected 45 variables, the option Heywood was used, which set the upper limit of communalities to 1. SAS Procedure Factor documentation (SAS Institute Inc., 2009) states that possible causes of communalities higher than 1 are: too many or too few common factors, not enough data, which affects the stability of estimations, problems with communalities estimations, or simply, the common factor model is not appropriate to explain the data. Moreover, high communalities seem to be a problem affecting both the 522-variable and the 45-variable input matrices. With this in mind, special care was taken in assessing the stability of the solutions across years. Finally, it was established that the range of possible interpretable factors varies between 7 and 5.

Factor analysis for the complete response rate matrix

The following discussion refers to the factor analysis of the complete response rate dataset (522 variables) for 2005, 2006, 2007 and 2008.

Examination of the tetrachoric correlation matrices showed that there are many variables with high correlation coefficients. These may indicate serious problems of heteroscedasticity. This is not surprising given the fact that we have more variables than observations. Nevertheless, PRINIT method works with singular correlation matrices. MSA (measure of sampling adequacy) could not be calculated for data entered in the form of a correlation matrix.

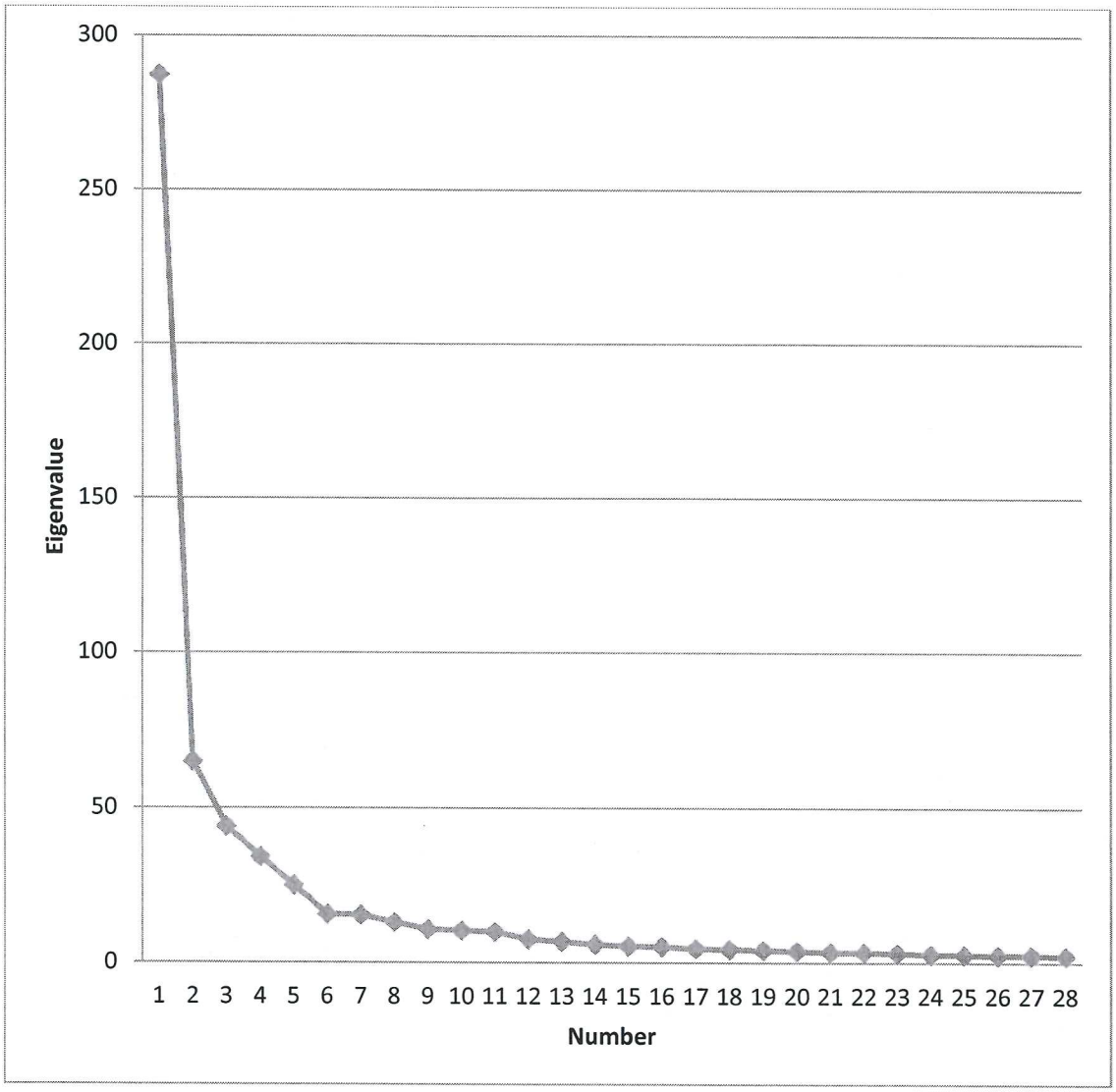
[4] From the set of 542 variables, around 20 presented problems of lack of variability (either zero or one for most countries) and were removed from the analysis.

For the decision of how many dimensions to retain, Larocque (2006) describes two "classic" methods. The first one, the criterion based on eigenvalues (or latent-root), stipulates that the number of factors must be equal to the number of eigenvalues greater than 1. Table 29 presents the eigenvalues related to the complete response rate matrix for 2007. It can be seen that the eigenvalues are very high, and indeed, the number of eigenvalues greater than one is 41 (not seen in Table 29), which is not suitable for data reduction purposes. The second method is called the scree test, and it is based in the visual inspection of the scree plot (Figure 15). The number of factors to retain in the scree test is the number of eigenvalues just before the beginning of the stabilization of the curve. Nevertheless, this number seems to be difficult to detect in the scree plot.

Table 29. Ten first eigenvalues for complete response rate matrix - 2007

Preliminary Eigenvalues: Total = 522 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	287.344548	222.377259	0.5505	0.5505
2	64.967289	20.916737	0.1245	0.6749
3	44.050551	9.757726	0.0844	0.7593
4	34.292825	9.151049	0.0657	0.8250
5	25.141776	9.434681	0.0482	0.8732
6	15.707095	0.279728	0.0301	0.9033
7	15.427367	2.353109	0.0296	0.9328
8	13.074259	2.254252	0.0250	0.9579
9	10.820007	0.413847	0.0207	0.9786
10	10.406159	0.374089	0.0199	0.9985

Figure 15. Scree Plot for Eigenvalues - Complete matrix data response rate for 2007



Stevens (2002) describes two additional methods: significance test of the retained number of components (which is affected by sample size) and the retention of factors accounting a minimum amount of the total variance (>70%). Due to problems with classic methods, the main approach used to choose the number of factors was interpretability of the dimensions, reproducibility in recent years and the retention of at least 70% of the variance. This last requirement is accomplished with a minimum of 3 factors.

The review of common factor analysis applied to the complete matrix of responses (522 variables) for years 2005, 2006, 2007 and 2008 - using both varimax (orthogonal - rotated factor pattern) and oblimin (oblique – rotated factor pattern, reference structure, factor structure) rotations- suggests a solution with 5 components. The 5-dimension solution is stable across

years and each dimension has a straightforward interpretation. In average, they represent 88% of the total variance between 2005 to 2008 data. Because the idea that factors related to the production of educational statistics are correlated seems plausible (financial data may come from a different source than data on enrolment but they may be generated by the same department or unit), the oblique rotation will be preferred through the analysis.

The proposed factors or dimensions are (including total, male, female and GPI where appropriate):

Factor 1 – Questionnaire A – detailed by grade enrolment/repeaters statistics of primary and secondary): enrolment in primary and secondary by grade/all grades, gross (and expected gross) intake ratio to the last grade of primary, expected gross primary graduation rate, gross intake ratio (GIR), GIR to the last grade of primary, new entrants to primary (grade 1, with previous ECCE programmes, total and female), repeaters in primary and secondary by grade/all grades, percentage of repeaters in primary and secondary by grade/all grades, repetition rate in primary by grade, survival rate to grade 4, 5 and last grade, transition from primary to secondary. This factor contains around 191 variables; these statistics are collected by UIS questionnaire A.

Factor 2 – Questionnaire A – general raw data enrolment/teaching statistics of primary and secondary: enrolment in primary, lower secondary, upper secondary, total secondary, post secondary non tertiary (public and private/private, by programme), teaching staff in primary, lower secondary, upper secondary, total secondary, post secondary (public and private, full and part-time, by programme), percentage of trained teachers in primary, lower, upper and total secondary, percentage of female teachers (primary, secondary), percentage of female students (primary, secondary), percentage of private enrolment (primary, secondary), pupil-teacher ratio (primary, secondary), technical/vocational enrolment as % of total enrolment (lower and upper secondary). This factor contains around 102 variables; these statistics are collected by UIS questionnaire A.

Factor 3 – Questionnaire A – net enrolment rate/gross enrolment rate/children out of school statistics of primary and secondary: adjusted net enrolment rate primary, gross enrolment rate (pre-primary, primary, lower secondary, upper secondary, secondary, primary and secondary combined), net enrolment ratio (pre-primary, primary and secondary), over-age and under-age enrolment rate, children out of school (rate primary school age, but in primary education), school life expectancy (years) (pre-primary, primary to secondary). This factor contains around 64 variables; these statistics are collected by UIS questionnaire A.

Factor 4 – Educational expenditure: all expenditure in education, such as: current expenditure as % of GNI, education expenditure by nature (ISCED 1 + 2 + 3 + 4; ISCED 5+6; capital, salaries, total), educational expenditure as % of total educational expenditure by ISCED, distribution (%) of public current expenditure by ISCED, public expenditure as % of GDP, GNI, total government expenditure, GDP per capita (by ISCED), distribution of total expenditure on educational institutions and administration as % of GDP (by source, by ISCED).This factor contains around 50 variables; these statistics are collected by UIS questionnaire B.

Factor 5 – Tertiary education statistics: all indicators and raw data on tertiary education, such as: distribution of students in ISCED 5A, 5B and 6, enrolment in 5A, 5B, and 6 (public and private), enrolment in all tertiary programmes, in total tertiary, gross completion rate ISCED 5A and 5 + 6, inbound mobility rate, number of students in tertiary education per 100,000 inhabitants, percentage of female students in ISCED 5A, 5B and 6, teaching staff in ISCED 5A, 5B and 6, total graduates, school life expectancy in tertiary. This factor contains around 50 variables; these statistics are collected by UIS questionnaire C.

The previous description did not include variables that have high loading in more than one factor. This issue will be addressed during the scale construction just after this discussion.

Table 30 and 31 show the correlation among factors (oblimin rotation) for 2007 and 2005 data, respectively. It can be seen that the coefficients vary from low to moderate (from 0.03 to 0.48).

Table 30. Correlation among factors (complete dataset - 2007)

Inter-Factor Correlations										
	Factor1		Factor2		Factor3		Factor4		Factor5	
Factor1 (QA)	100	*	48	*	33	*	10		37	*
Factor2 (QA)	48	*	100	*	20		3		25	
Factor3 (QA)	33	*	20		100	*	13		23	
Factor4 (QB)	10		3		13		100	*	32	*
Factor5 (QC)	37	*	25		23		32	*	100	*
Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.3 are flagged by an '*'.										

Table 31. Correlation among factors (oblimin rotation with 5 factors - 2005)

Inter-Factor Correlations										
	Factor1		Factor2		Factor3		Factor4		Factor5	
Factor1 (QA1)	100	*	47	*	44	*	11		34	*
Factor2 (QA2)	47	*	100	*	30	*	4		20	
Factor3 (QA3)	44	*	30	*	100	*	18		29	
Factor4 (QB)	11		4		18		100	*	30	*
Factor5 (QC)	34	*	20		29		30	*	100	*
Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.3 are flagged by an '*'.										

A partial view of the residual correlations with uniqueness on the diagonal for 2007 after oblimin rotation is presented in Table 32. It can be noted that many elements off the diagonal are small, although there are some residual correlations between 0.1 and 0.2. In addition, the overall root mean square off-diagonal residual is 0.083, and communalities are close to 1, indicating a model with a possible good fit. Something important to remark is the fact that, although communalities are high, the partial correlation controlling for factors present many elements greater than 1, being the overall root mean square off-diagonal partials equal to 2.7 (the maximum should be 1). As it will be shown next, part of the problem with the partial correlations controlling for factors is due to presence of negative eigenvalues in the initial correlation matrix.

Table 32. Partial view of residual correlations with uniqueness on the diagonal (Oblimin rotation) - 2007

Residual Correlations With Uniqueness on the Diagonal										
	PF5AT	PF5BT	PF6T	PFSAG	PFSAP	PFSAT	PFSI0	PFSI1	PFSI4	PFTTT
PF5AT	-3	0	9	0	-4	0	-5	-3	0	-3
PF5BT	0	2	10	-1	-6	1	-7	2	-2	-4
PF6T	9	10	27	-7	-14	-4	-20	-6	-11	2
PFSAG	0	-1	-7	-1	-4	5	-2	-10	-8	-1
PFSAP	-4	-6	-14	-4	-3	4	3	-2	2	-2
PFSAT	0	1	-4	5	4	22	0	-2	6	-2
PFSI0	-5	-7	-20	-2	3	0	28	5	-3	-6
PFSI1	-3	2	-6	-10	-2	-2	5	-15	-7	-5
PFSI4	0	-2	-11	-8	2	6	-3	-7	41	1
PFTTT	-3	-4	2	-1	-2	-2	-6	-5	1	-3

Printed values are multiplied by 100 and rounded to the nearest integer

Factor analysis for the 45 selected variables and scale building

Even though the suggested 5-dimension structure is a robust solution across years, the problems of the item to observation ratio and the high partial correlations needed be addressed. To do that, a second step was carried. Factor analysis was applied to the group of 45 variables described in section 7.4, and compared to the previous 5-dimension solution. Then, basic principles of the construction of questionnaires were applied (e.g. examination of loadings, deletion of variables with loading below 0.5 or with high loading in more than one dimension, addition of variables with high loading in one specific dimension in order to obtain a 45-item scale with 5 dimensions also). This scale is employed successfully in the exploration of other aspects of the education database in subsequent sections.

Factor analysis in the group of 45 variables also indicates a solution of 5 dimensions, but further improvements were needed in order to obtain a reliable scale. From this original group, nine variables were retired from analysis due to problems with loadings (high in more than one dimension or low loadings in all dimensions). Most of the time, the retired variable was replaced by another variable, similar in interpretation and importance to the remaining variables (reasonable as proxy measure) and with high loading in the respective factor. There were 18 variables with high loading in Factor 1, but six variables that had high loading in also other

factors were removed and not replaced in order to keep the balance of items across factor. A similar number of variables were distributed among the other factors for the same reason (balance of items across factors).

The number of items per dimension of the 45-item scale is as follows (also shown in Table 43):

- Factor 1 (QA): 12 variables;
- Factor 2 (QA): 8 variables;
- Factor 3 (QA): 10 variables;
- Factor 4 (QB): 8 variables; and
- Factor 5 (QC): 7 variables.

The basic principles of model examination were applied as in the case of the complete response rate matrix. Across years (1999 to 2008), the retained 5 dimensions represent 90% of the variability. The communalities are generally high, the residuals correlations also low, and as seen in the standardized regression coefficient for 2007 data (shown in Table 33), most of the times loading are high in the respective dimension and low in others. The analysis of the rotated factor pattern, the reference structure and the factor structure matrices give identical results within each year, and across years. The results are very similar, with only few occasions where variables are not loading high in the hypothesized dimension. For example, this is the case of item PEPTF, which theoretically belong to Factor 2, but for 2007 it is loading high in Factor 3. For other years, PEPTF is loading in the hypothesized factor (factor 2–QA2). The correlations among factors (shown in Table 34) are similar to those related to the 5–dimension models with the complete dataset.

Table 33. Rotated factor pattern for the 45-item scale (2007)

Rotated Factor Pattern (Standardized Regression Coefficients)							
Item Code	Item Name	Factor1	Factor2	Factor3	Factor4	Factor5	Hypothesized Factor
GPTR	Gender parity index for transition rate, primary to secondary, general programmes	95	5	-16	13	3	Factor1 (QA)
TRANT	Transition from ISCED 1 to ISCED 2, general programmes (%). Total	93	5	-16	10	11	Factor1 (QA)
SR5FF	Survival rate to grade 5. Female	85	-5	-5	9	8	Factor1 (QA)
PRFF	Percentage of repeaters in primary. All grades. Female	83	4	19	2	4	Factor1 (QA)
SR5FT	Survival rate to grade 5. Total	83	-1	-6	11	9	Factor1 (QA)
GPGIL	Gender parity index for gross intake ratio to the last grade of primary	82	2	25	-3	-4	Factor1 (QA)
PRFT	Percentage of repeaters in primary. All grades. Total	81	12	18	-4	13	Factor1 (QA)
GIRLT	Gross intake ratio to the last grade of primary. Total	78	5	27	-3	1	Factor1 (QA)
GPAIR	Gender parity index for gross intake ratio. Primary	71	8	35	0	-8	Factor1 (QA)
AIRFT	Gross intake ratio. Primary. Total	67	11	35	-1	0	Factor1 (QA)
PRST	Percentage of repeaters in secondary. All grades. Total	50	16	31	-3	16	Factor1 (QA)
PRSF	Percentage of repeaters in secondary. All grades. Female	50	20	27	4	8	Factor1 (QA)
TRASF	Percentage of trained teachers. Total secondary. Female	-13	103	-3	-11	7	Factor2 (QA)
TRAST	Percentage of trained teachers. Secondary (ISCED 2 and 3). Total	-10	96	8	-12	1	Factor2 (QA)
TRA1T	Percentage of trained teachers. Primary. Total	15	92	-4	-13	-13	Factor2 (QA)
TRA1F	Percentage of trained teachers. Primary. Female	13	89	-9	-8	-13	Factor2 (QA)
PTRF	Pupil-teacher ratio. Primary	29	79	-2	23	1	Factor2 (QA)
PTRS	Pupil-teacher ratio. Secondary	-7	79	22	22	14	Factor2 (QA)
PEPTF	Percentage of private enrolment. Primary	27	44	25	41	1	Factor2 (QA)
GPNES	Gender parity index for net enrolment rate. Secondary	-18	4	102	6	3	Factor3 (QA)
NERST	Net enrolment rate. Secondary. All programmes. Total	-13	3	97	6	5	Factor3 (QA)
ROFT	Rate of primary school age children out of school. Total	23	-5	81	3	3	Factor3 (QA)
NERFT	Net enrolment rate. Primary. Total	23	-5	81	3	3	Factor3 (QA)

Rotated Factor Pattern (Standardized Regression Coefficients)							
Item Code	Item Name	Factor1	Factor2	Factor3	Factor4	Factor5	Hypothesized Factor
ROFF	Rate of primary school age children out of school. Female	19	-4	80	-2	6	Factor3 (QA)
GPNEP	Gender parity index for net enrolment rate. Primary	26	-5	79	1	1	Factor3 (QA)
GERFT	Gross enrolment ratio. Primary. Total	37	16	71	11	-3	Factor3 (QA)
GERST	Gross enrolment ratio. Secondary. All programmes. Total	5	29	68	10	17	Factor3 (QA)
GPGES	Gender parity index for gross enrolment ratio. Secondary. All programmes	3	33	65	15	12	Factor3 (QA)
GPGEP	Gender parity index for gross enrolment ratio. Primary	44	11	63	15	-7	Factor3 (QA)
TVTSP	Technical/vocational enrolment in ISCED 2 and 3 as % of total enrolment in ISCED 2 and 3	-6	28	46	16	29	Factor2 (QA)
ECSTO	Percentage distribution of public current expenditure on education by level. Secondary	-7	1	0	105	-6	Factor4 (QB)
EC1TO	Percentage distribution of public current expenditure on education by level. Primary	-3	-2	3	102	-7	Factor4 (QB)
ECTTO	Percentage distribution of public current expenditure on education by level. Tertiary	-5	-4	-3	101	1	Factor4 (QB)
ECNTO	Percentage distribution of public current expenditure on education not allocated by level	4	0	-5	98	1	Factor4 (QB)
EC0TO	Percentage distribution of public current expenditure on education by level. Pre-primary	14	-13	0	94	-2	Factor4 (QB)
EEGDP	Public expenditure on education as % of GDP	8	-1	0	91	4	Factor4 (QB)
PCGDP	Public expenditure per pupil as a % of GDP per capita. Primary	8	-7	3	86	2	Factor4 (QB)
EEGE	Public expenditure on education as % of total government expenditure	-3	7	5	82	16	Factor4 (QB)
R25003	Teaching staff in total tertiary. Public and private. Full and part-time. All programmes. Total	12	9	-16	-8	100	Factor5 (QC)
R25007	Teaching staff in total tertiary. Public and private. Full and part-time. All programmes. Female	6	8	-9	-8	99	Factor5 (QC)
E26375	Total graduates in all programmes. Tertiary. Total	1	-13	4	6	93	Factor5 (QC)
E26415	Total graduates in all programmes. Tertiary. Female	2	-10	0	11	91	Factor5 (QC)
GERTF	Gross enrolment ratio. ISCED 5 and 6. Female	10	-10	12	-2	90	Factor5 (QC)
GERTT	Gross enrolment ratio. ISCED 5 and 6. Total	8	-9	16	-2	87	Factor5 (QC)

Rotated Factor Pattern (Standardized Regression Coefficients)							
Item Code	Item Name	Factor1	Factor2	Factor3	Factor4	Factor5	Hypothesized Factor
FSPTE	Inbound mobility rate. Total	-21	6	6	13	85	Factor5 (QC)
Printed values are multiplied by 100 and rounded to the nearest integer.							

Table 34. Correlations among factors (45-item scale) for 2007

Inter-Factor Correlations										
	Factor1		Factor2		Factor3		Factor4		Factor5	
Factor1	100	*	34	*	50	*	36	*	28	
Factor2	34	*	100	*	39	*	0		11	
Factor3	50	*	39	*	100	*	37	*	34	*
Factor4	36	*	0		37	*	100	*	40	*
Factor5	28		11		34	*	40	*	100	*
Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.3 are flagged by an '*'.										

The 45-item scale represents a ratio subject to item of 4.6:1, which is close to the recommended ratio of 5:1. Nevertheless, the matrix of partial correlations controlling by factors, as in the case of the factor analysis of the complete dataset, has values greater or closer to 1. Table 35 presents a partial view of the partial correlation controlling by factors for the 45–item scale (5–dimension solution) for 2007, where these problems become evident (many partial correlations over 1). The overall root mean square off-diagonal partials for this model is 0.6155 (the maximum is 1). For a well fitted model, this value is expected to be close to 0.

Table 35. Partial view of partial correlation controlling factors - correlation matrix from tetrachoric correlation matrix - 45-item scale - 2007

	Partial Correlations Controlling Factors								
	GERFT	GERST	GERTF	GERTT	GPGEP	GPGES	GPNEP	GPNES	NERFT
GERFT	0	0	0	0	0	0	0	0	0
GERST	0	100	68	15	0	112	-132	299	-137
GERTF	0	68	100	104	0	71	-20	-49	-17
GERTT	0	15	104	100	0	36	11	-133	13
GPGEP	0	0	0	0	0	0	0	0	0
GPGES	0	112	71	36	0	100	-78	65	-75
GPNEP	0	-132	-20	11	0	-78	100	-118	102
GPNES	0	299	-49	-133	0	65	-118	100	-134
NERFT	0	-137	-17	13	0	-75	102	-134	100

Note: printed values are multiplied by 100 and rounded to the nearest integer.

One possible cause of this problem is the presence of negative eigenvalues in the tetrachoric correlation matrix used in the factor analysis. For comparison reasons, a factor analysis was applied on the 45–item scale, but directly on the binary data (no calculation of the tetrachoric correlation matrix). The results were similar to the solution find through the factor analysis on the tetrachoric correlation matrix. Nevertheless, the partial correlations controlling by factors were all below 1, although some values are well over 0.4, which are seemingly high for a well fitted model (see Table 36). The overall root mean square off-diagonal partials for this model is 0.1726, value closer to 0 than that of the model with tetrachoric correlation matrix (0.6155).

Table 36. Partial view of the Partial correlation controlling factors – correlation matrix from binary data input – 45-item scale – 2007

Partial Correlations Controlling Factors									
	GERFT	GERST	GERTF	GERTT	GPGEP	GPGES	GPNEP	GPNES	NERFT
GERFT	100	43	12	18	94	31	-34	-25	-34
GERST	43	100	7	2	41	84	-44	0	-46
GERTF	12	7	100	81	12	9	-6	-4	-6
GERTT	18	2	81	100	17	5	-2	-12	-2
GPGEP	94	41	12	17	100	29	-28	-27	-27
GPGES	31	84	9	5	29	100	-41	5	-42
GPNEP	-34	-44	-6	-2	-28	-41	100	-14	81
GPNES	-25	0	-4	-12	-27	5	-14	100	-21
NERFT	-34	-46	-6	-2	-27	-42	81	-21	100

Note: Printed values are multiplied by 100 and rounded to the nearest integer.

For further comparison, the tetrachoric correlation matrix for the 45-item scale was smoothed (removal of the components related to negative eigenvalues) using a software called TetMat (Uebersax, 2007) and factor analysis was applied on it. A solution with 5–dimensions seemed also appropriate, and the results, including loadings, are very similar with the obtained through factor analysis on the tetrachoric correlation matrix and with the binary data. As shown in Table 37, the partial correlations for the model related to the smoothed tetrachoric correlation matrix are all below 1 and closer to the values obtained through the factor analysis on binary data. The overall root mean square off-diagonal partials for this model is 0.2699, value closer to 0 than that of the model with tetrachoric correlation matrix (0.6155) but a little higher than that from the model on binary data (0.1726).

Table 37. Partial view of the Partial correlation controlling factors – smoothed tetrachoric correlation matrix – 45-item scale – 2007

Partial Correlations Controlling Factors									
	GERFT	GERST	GERTF	GERTT	GPGEP	GPGES	GPNEP	GPNES	NERFT
GERFT	100	20	47	38	33	12	-26	13	-21
GERST	20	100	16	-1	14	75	-60	55	-66
GERTF	47	16	100	78	50	22	-8	-6	-10
GERTT	38	-1	78	100	64	18	9	-35	13
GPGEP	33	14	50	64	100	16	20	-45	28
GPGES	12	75	22	18	16	100	-39	23	-50
GPNEP	-26	-60	-8	9	20	-39	100	-49	88
GPNES	13	55	-6	-35	-45	23	-49	100	-56
NERFT	-21	-66	-10	13	28	-50	88	-56	100

Printed values are multiplied by 100 and rounded to the nearest integer.

As a reference, Table 38 presents the standardized regression coefficients (Rotated Factor Pattern) for Factor1 on the 45–item scale for 2007 obtained after factor analysis on the tetrachoric correlation matrix, the smoothed tetrachoric matrix and the binary data (oblimin rotation). The factor loadings on a model applied directly on binary data seems to be underestimated [situation also pointed by Wood, C.M. (2002)]. The coefficients related to factor analysis on the smoothed tetrachoric matrix are greater than those from the model on binary data but lower to those obtained through factor analysis on the tetrachoric correlation matrix. These behaviours are observed across all factors.

Table 38. Standardized regression coefficients (Rotated Factor Pattern) for Factor1-QA1 on the 45-item scale for 2007

Item Code	Item Name - Factor 1-QA1	Tetrachoric correlation matrix 2007	Smoothed tetrachoric matrix 2007	Binary data 2007
GPTR	Gender parity index for transition rate, primary to secondary, general programmes	95	93	83
GPGIL	Gender parity index for gross intake ratio to the last grade of primary	82	79	78
GPAIR	Gender parity index for gross intake ratio. Primary	71	69	63
GIRLT	Gross intake ratio to the last grade of primary. Total	78	76	75
AIRFT	Gross intake ratio. Primary. Total	67	65	61
PRFF	Percentage of repeaters in primary. All grades. Female	83	80	72
PRFT	Percentage of repeaters in primary. All grades. Total	81	78	70
PRSF	Percentage of repeaters in secondary. All grades. Female	50	49	41
PRST	Percentage of repeaters in secondary. All grades. Total	50	49	41
SR5FF	Survival rate to grade 5. Female	85	85	70
SR5FT	Survival rate to grade 5. Total	83	83	70
TRANT	Transition from ISCED 1 to ISCED 2, general programmes (%). Total	93	91	82

In short, we can conclude that the 5–dimension model is a good representation of the structure of the response rates in the education database. Moreover, the 45–item scale seems to be a reasonable approximation to the response rate structure of the complete set of variables.

CHAPTER 7. Classification of countries

After the analysis of the structure of variable responses (factor analysis in Chapter 6), cluster analysis was applied for the classification of countries - based on their patterns of response - each year from 1999 to 2009. First, information from the selected 45 indicators on each country (see section 6.3 – *Factor analysis for the 45 selected variables and scale building*) was used directly in a clustering algorithm. Afterwards, the countries' factor scores[5] on each of the five dimensions were used to obtain a meaningful description of the clusters. The proposed classification of countries (5 clusters or subgroups) is consistent across years, and it improves the understanding of the response behaviour of countries around the world.

7.1 Objectives

The main objectives of this chapter are to classify countries based on the capacity to respond to the 45 items previously proposed and to present an interpretation that can be useful for future diagnosis.

7.2 Statistical model

The first step is to transform the dataset of 45 items – selected during the scale's development in the previous section – into a matrix of distances (dissimilitude). For this objective, the Dmatch method from SAS' procedure distance (SAS Institute Inc., 2009) was used, which converts the simply matching coefficient – a type of association measure – to a Euclidian distance. This type of dissimilitude measure for binary variables is widely recommended (Larocque, 2006). Then, the resulting matrix of distances was used as input for the cluster analysis – a hierarchical clustering procedure with Ward's method. This procedure was carried with data for each year (10 times, from 1999 to 2008). It is worth to remark that the factor scores could have been used directly into the clustering algorithm; nevertheless, the proposed factors have similar number of items, therefore, using factor scores or using the information of the 45 items directly would give similar results.

7.3 Cluster analysis: results and comments

From the analysis of the dendograms (see Figures 16 and 17), it can be seen that a reasonable choice for the number of clusters falls in the range of 3 to 7 clusters. The examination of results with different number of clusters, while taking into account the need for a small number of

[5] Taking into account the 45 selected variables, a factor score is the average of available responses related to variables belonging to the factor divided by the total number of variables in the respective factor

groups that can help with the efficient description of countries, pointed to the choice of 5 clusters each year.

The interpretation of the 5–cluster solution using the proposed scale is immediate for years 1999, 2002, 2004, 2005, 2006, 2007 and 2008. For years 2000, 2001 and 2003, some clusters are more difficult to describe and were not taken into consideration in the present analysis. The cluster mean values on each factor are presented in Table 39. Figures 18 to 22 present the same values.

Figure 16. Cluster analysis - Dendrogram from Ward's method - 2007 (45 selected items)

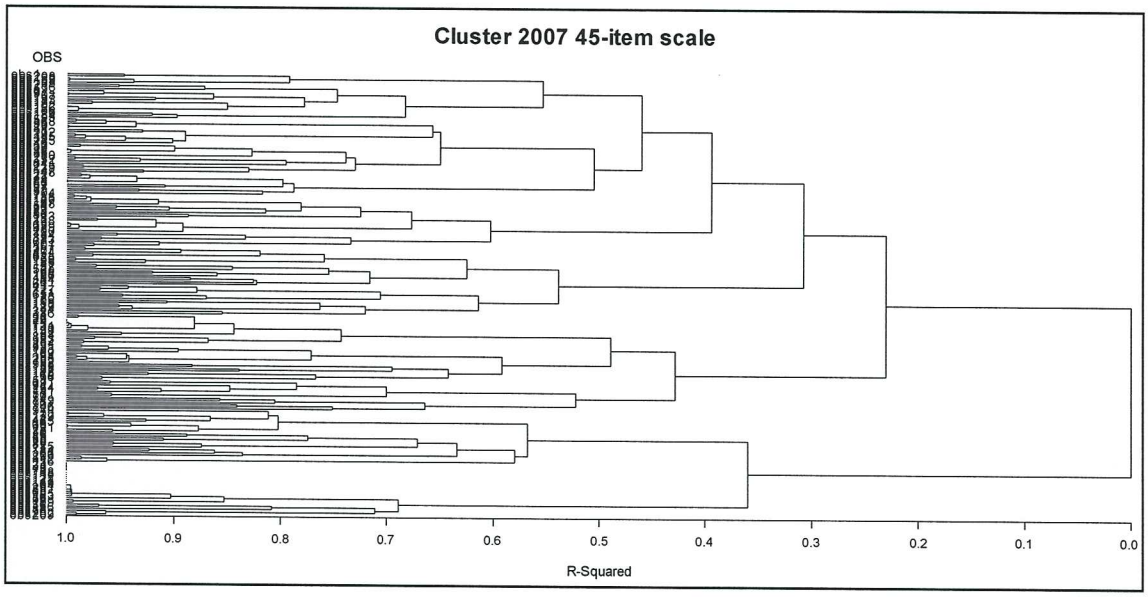


Figure 17. Cluster analysis - Dendrogram from Ward's method - 2004 (45 selected items)

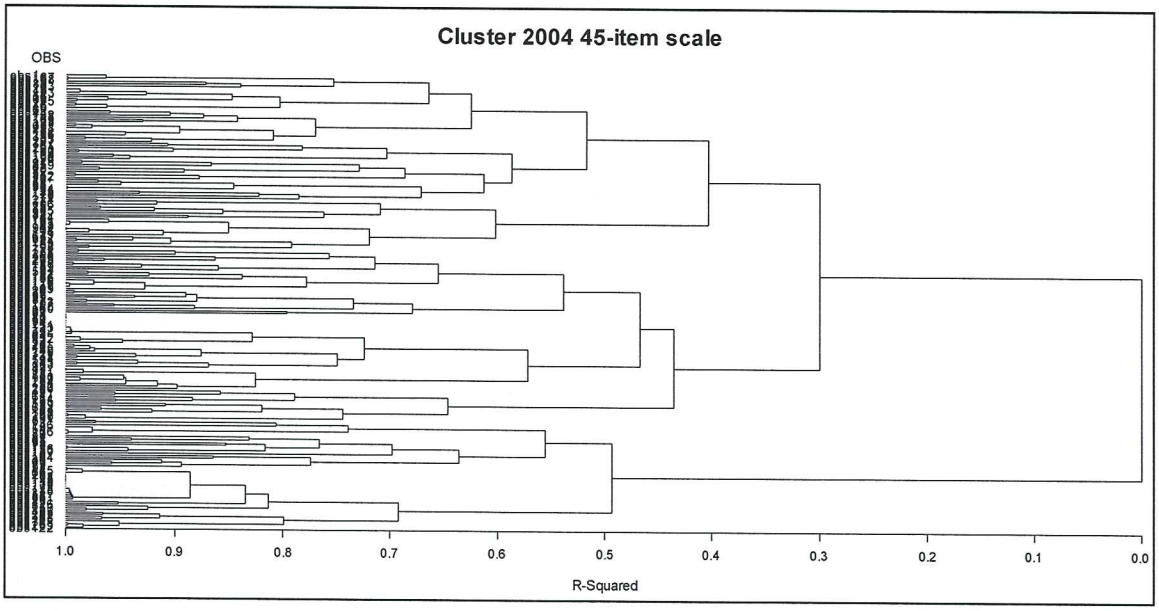


Table 39. Cluster response rate means by factor

Year	Cluster	Frequency	Factor 1 (QA)	Factor 2 (QA)	Factor 3 (QA)	Factor 4 (QB)	Factor 5 (QC)
1999	1	45	0.689	0.517	0.782	0.928	0.638
	2	82	0.911	0.593	0.848	0.226	0.566
	3	32	0.401	0.398	0.784	0.160	0.549
	4	25	0.070	0.040	0.048	0.025	0.034
	5	25	0.213	0.290	0.192	0.115	0.754
2002	1	75	0.856	0.573	0.872	0.868	0.695
	2	36	0.741	0.514	0.489	0.111	0.540
	3	35	0.874	0.568	0.940	0.139	0.224
	4	40	0.096	0.122	0.168	0.084	0.336
	5	23	0.475	0.505	0.970	0.239	0.901
2004	1	65	0.901	0.573	0.852	0.952	0.697
	2	57	0.870	0.572	0.805	0.169	0.742
	3	23	0.797	0.712	0.948	0.071	0.112
	4	50	0.090	0.138	0.150	0.100	0.274
	5	14	0.345	0.464	0.943	0.714	0.898
2005	1	48	0.793	0.542	0.988	0.919	0.869
	2	65	0.796	0.529	0.834	0.098	0.679
	3	23	0.804	0.707	0.791	0.125	0.025
	4	44	0.119	0.111	0.148	0.060	0.146
	5	29	0.750	0.634	0.717	0.918	0.443
2006	1	52	0.889	0.541	0.873	0.947	0.701
	2	51	0.824	0.618	0.914	0.201	0.843
	3	39	0.774	0.571	0.733	0.090	0.070
	4	54	0.090	0.113	0.065	0.076	0.235
	5	13	0.186	0.481	0.854	0.702	0.626
2007	1	65	0.827	0.548	0.888	0.933	0.692
	2	50	0.787	0.595	0.888	0.168	0.877
	3	37	0.743	0.554	0.862	0.071	0.069
	4	33	0.073	0.121	0.006	0.114	0.255
	5	24	0.462	0.453	0.433	0.068	0.107
2008	1	62	0.790	0.587	0.876	0.833	0.758
	2	44	0.739	0.619	0.900	0.099	0.821
	3	22	0.777	0.722	0.918	0.068	0.058
	4	46	0.020	0.103	0.067	0.082	0.273
	5	35	0.695	0.436	0.517	0.421	0.159

As mentioned before, the scores of response on each of the five dimensions were used to understand the 5-cluster solutions.

The interpretation of the five clusters is as follows:

- Cluster 1: This cluster is composed by countries that have high response rates (group average per year over 0.7) in Factors 1, 3, 4 and 5. Factor 2 has an average response rate

per year varying from 0.5 to 0.6, which can be considered high, but that is comparatively lower than the average response rate for other factors (see Figure 18). In average, this group includes 59 countries (representing 28% of the total of 209).

- Cluster 2: This cluster is similar to cluster 1, the main difference being that countries in cluster 2 have low response rate for Factor 4 (less than 0.20 across years), which the dimension is related to the response of education finance's indicators (see Figure 19). In average, 55 countries (26%) are found in this group.
- Cluster 3: This cluster is composed by countries that have relative high response rates (but slightly lower than in Cluster 1) for Factor 1, 2 and 3, but that exhibit very low response rates in Factor 4 (education finances) and 5 (tertiary education statistics) (less than 0.15 across years). Factor 5 is the dimension related to the response of tertiary education's indicators (see Figure 20). In average, 30 countries (30%) are found in this group.
- Cluster 4: This cluster is composed by countries that have very low response rates in all factors (lower than 0.2, except for Factor 5, which is in average less than 0.3) (see Figure 21). In average, 41 countries (20%) are found in this group.
- Cluster 5: This cluster is the most difficult to describe; it seems to be made of the remaining of countries that do not fit well in any of the previous clusters each year (see Figure 22). In average, 23 countries (11%) are found in this group.

The clusters' numeration reflects a preferable behaviour of response. For example: countries in Cluster 1 perform well in all factors, while countries in Cluster 2 perform well in all dimensions except in Factor 4 and countries in Cluster 3 perform relatively well only in the first three factors and have poor performances in Factors 4 and 5; at the same time, countries in cluster 4 perform poorly in all factors. Countries in Cluster 5 could be considered as non comparable as it may include the remaining countries that do not fit in other clusters.

Figure 18. Evolution of the average response rate by factor - CLUSTER 1

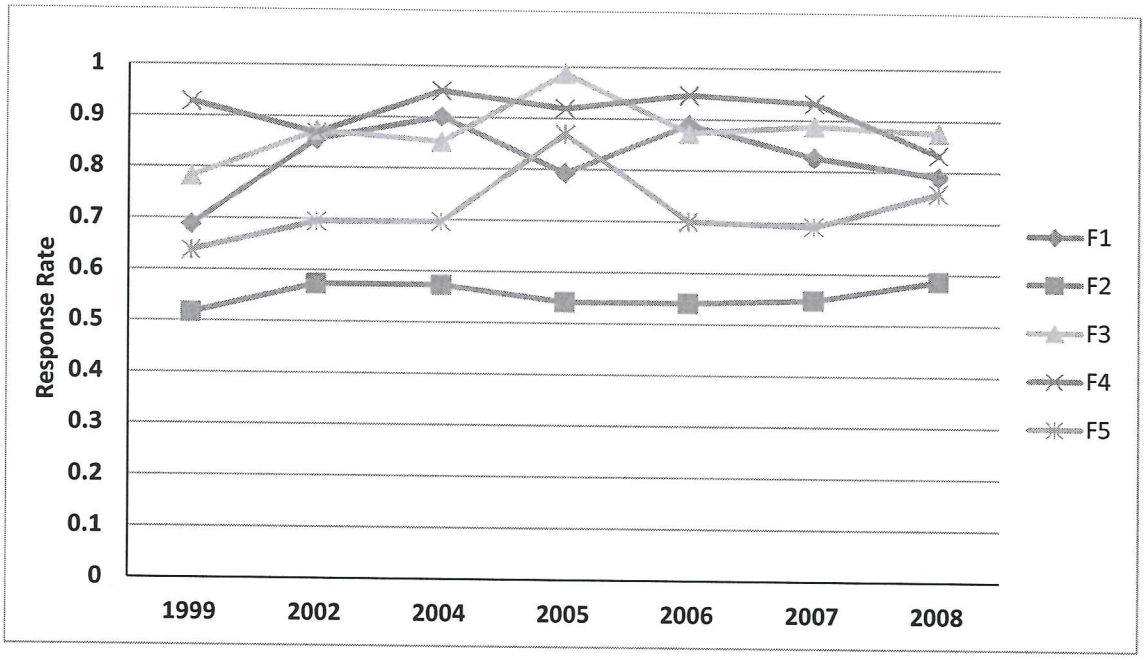


Figure 19. Evolution of the average response rate by factor – CLUSTER 2

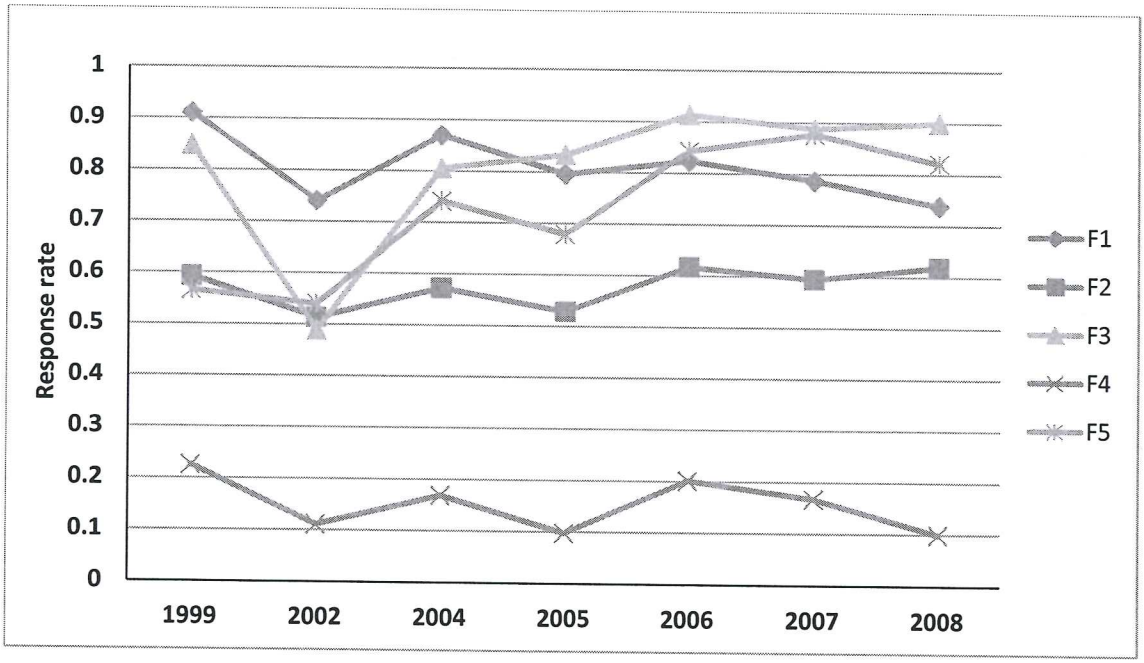


Figure 20. Evolution of the average response rate by factor – CLUSTER 3

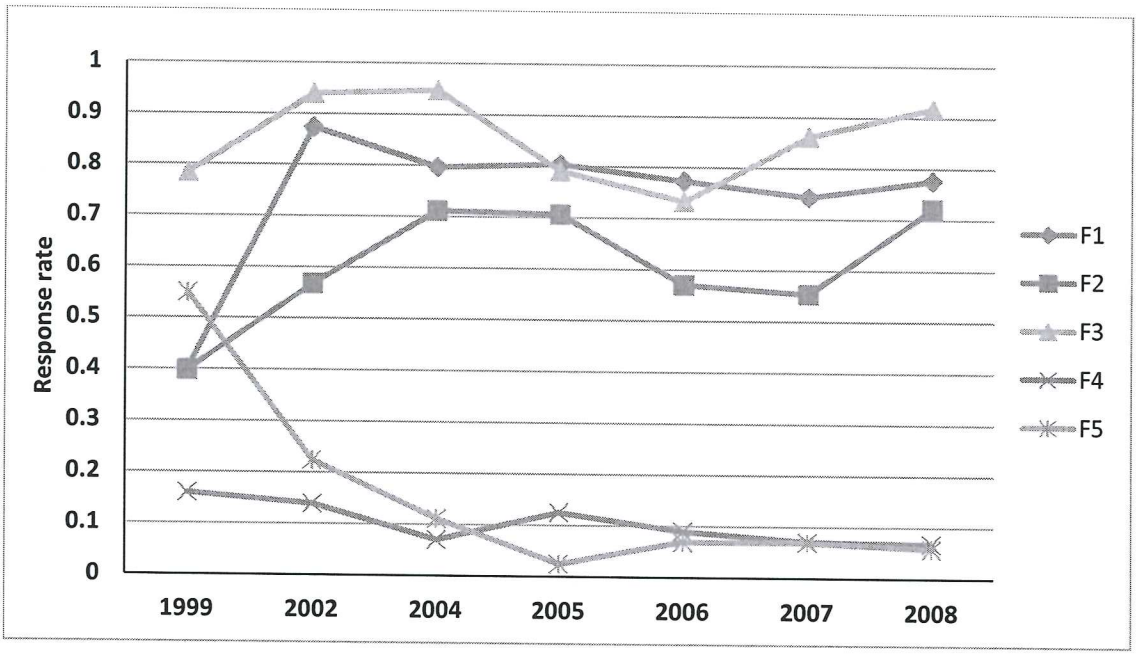


Figure 21. Evolution of the average response rate by factor – CLUSTER 4

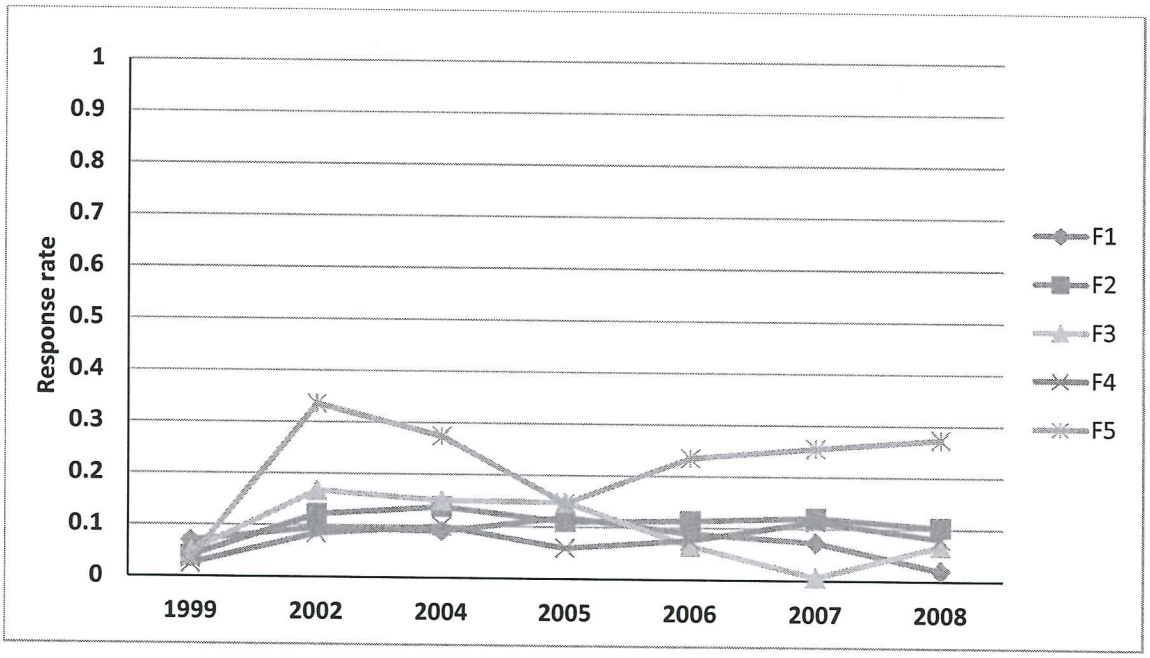
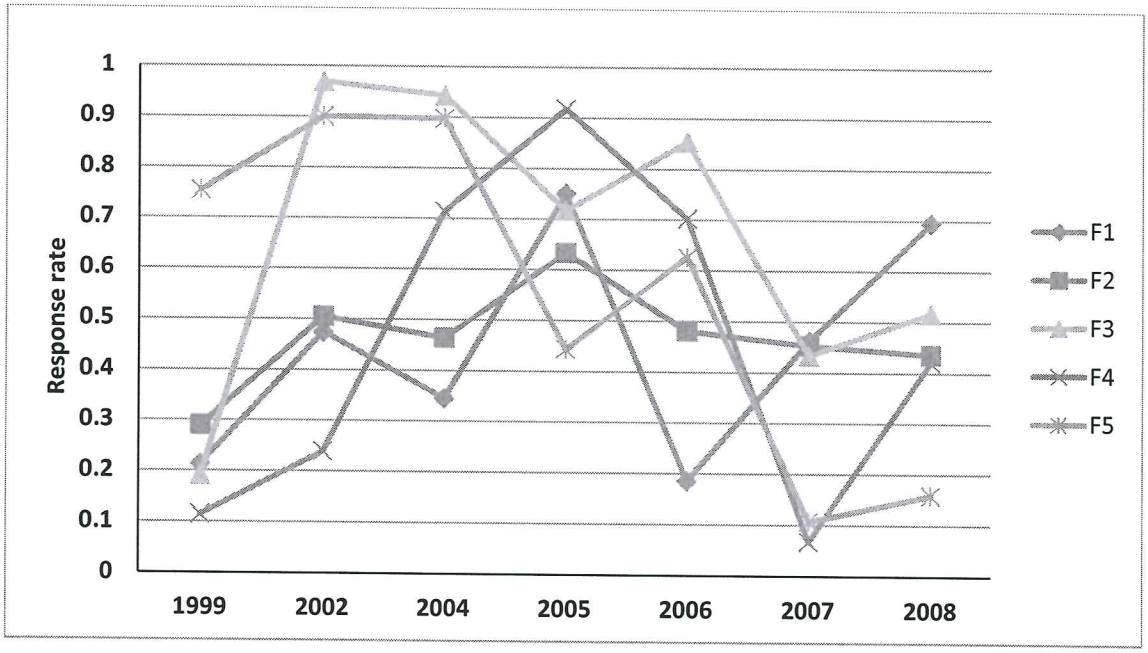


Figure 22. Evolution of the average response rate by factor – CLUSTER 5



We can note from Figure 18 to 22 that the suggested classifications, from Cluster 1 to Cluster 4, are stable in time. This is important to remark as it gives support to the validity of the proposed solution.

Another important fact is that, for the best performing countries (Cluster1), items from Factor 2 are in some degree difficult to report (its response rate trend is lower than the trend of other dimensions) (see Figure 18), although the average response rate to items seem stable across years. This dimension represents general data about "enrolment and teaching staff" statistics of primary and secondary, such as the ratio pupil/teachers for primary and secondary. It also seems that items from Factor 2 are relatively difficult to respond for the rest of clusters in comparison, at least, to items from Factor 1 and 3 (items from Factor 1, 2 and 3 are collected by the same education questionnaire, which is related to statistics for pre-primary, primary, secondary and post-secondary non tertiary).

Do countries find useful to produce this type of indicators (Factor 2) at the national level? Are countries using other type of statistics that measure similar concepts as the UIS indicators encompassed in dimension Factor 2? Or is that the UIS questionnaire is not well designed to collect these statistics at the national level? These questions are pertinent to understand the possible problems with the collection and production of items related to Factor 2.

Another important aspect to remark is the usefulness of the five factors or dimensions proposed for the 45-item scale in the interpretation of the clusters.

Respect to the cluster membership, we can also expect that countries that are placed in certain cluster one year can be found in the same cluster in other years. Table 40 presents the list of countries that were classified in the respective cluster at least 5 times in the analysis of years 1999, 2002, 2004-2008 (5 or more times out of 7 years). In total, we find that 95 countries fulfil this condition. It can be noted that the most stable groups are Cluster 1 (40 constant members out of 59 members on average) and Cluster 4 (27 constant member out of 41 member on average). Cluster 2 has 21 constant members out of 55 in average and Cluster 3 has 7 constant members out of 30 in average.

Table 40. Countries with constant membership in a given cluster (5 times or more in 7 years)

Cluster	Number of countries	Countries (5 or more times as members of the cluster - 1999, 2002, 2004-2008)
Cluster 1	**40**	**Argentina**, Aruba, **Australia**, Austria, **Azerbaijan**, Bulgaria, **Hong Kong**, Colombia, **Cuba**, Cyprus, **Czech Republic**, Denmark, **El Salvador**, Estonia, **Finland**, France, **Hungary**, Iceland, **Iran** , Ireland, **Israel**, Italy, **Lesotho**, Lithuania, **Madagascar**, Mali, **Mauritius**, Mexico, **New Zealand**, Norway, **Philippines**, Poland, **Republic of Korea**, Slovakia, **South Africa**, Spain, **Sweden**, Switzerland, **Togo**, Tunisia.
Cluster 2	**21**	**Algeria**, Brunei Darussalam, **Macao**, Ethiopia, **Georgia**, Jordan, **Kazakhstan**, Lao, **Latvia**, Malawi, **Montserrat**, TO Palestine, **Pakistan**, Panama, **Russian Federation**, Sao Tome and Principe, **Tajikistan**, FYR Macedonia, **Ukraine**, Uruguay, **Uzbekistan**.
Cluster 3	**7**	**Bahamas**, Dominican Republic, **Gambia**, Holy See, **Myanmar**, Nicaragua, **United Arab Emirates**.
Cluster 4	**27**	**Afghanistan**, Angola, **Antigua and Barbuda**, Bosnia and Herzegovina, **DPR Korea**, Gabon, **Gibraltar**, Guinea-Bissau, **Haiti**, Libya, **Micronesia**, Monaco, **Montenegro**, Netherlands Antilles, **Oman**, Palau, **Papua New Guinea**, Puerto Rico, **San Marino**, Sierra Leone, **Singapore**, Somalia, **Timor-Leste**, Tokelau, **Turkmenistan**, Viet Nam, **Zimbabwe**.

Note: Text in bold was used to improve readability.

Countries were classified into a cluster each year. If we consider clusters as states, this classification may represent a country moving from one state to another. In that case, we could construct a country average – all countries confound – matrix of transition probability. Table 41 presents the probabilities of passing from a given cluster (year t) one year to another cluster (year t+1) – including resting in the same cluster – the following year for the "average country".

The probabilities were calculated only from 2004 to 2008 (2004 is not consecutive to 2002; therefore, 2002 data were excluded).

Table 41. Matrix of transition probabilities for Cluster 1 to 5 – 2004 to 2008

		Year t+1				
		Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Year t	**Cluster 1**	**0.609**	0.174	0.061	0.013	0.143
	Cluster 2	0.166	**0.556**	0.130	0.076	0.072
	Cluster 3	0.090	0.156	**0.443**	0.131	0.180
	Cluster 4	0.039	0.083	0.094	**0.674**	0.110
	Cluster 5	0.400	0.150	0.088	0.238	**0.125**

With exception of countries in Clusters 3 and 5, it can be noted that given their current clusters most countries have at least 50% probabilities of being classified in the same cluster again for a consecutive year, with countries in Clusters 1 and 4 displaying the highest probabilities (61% and 67.4% respectively). For countries in Cluster 3, the probability is 44%. The high probabilities of remaining in the same cluster for countries in Clusters 1 and 4 are not surprising: countries in Cluster 1, the best performing cluster, can be considered as having the best education statistical capacities in term of high response rates, while for countries in Cluster 4, the opposite is true. In other words, countries in Cluster 1 exhibit consistency in their membership to Cluster 1 due to a robust education statistical capacity, while countries in Cluster 4 cannot improve their response rates permanently as their statistical systems may not have the necessary capacity.

Regarding the countries' transition probability of moving from a cluster to a different one, we can note that, excluding Cluster 5, these probabilities decrease as the distance between original and destiny clusters increase. For example, the probabilities of a country classified in Cluster 1 a given year to be classified in Clusters 3 or 4 (6% and 1%, respectively) the following year are less than the probability of being classified in Cluster 2 (17%). For countries in Cluster 1, 2, 3 and 4, the transition probabilities of changing cluster classification are in general less than 18% for any new cluster.

We can also note that countries in Cluster 5 have high probabilities of being classified in different clusters. This is expected as countries in Cluster 5 seem to be the countries that do not fit into Cluster 1 to 4 in a given year.

Table 42 presents countries placed a maximum of 3 times (out of 7 years) in a given cluster. It can be noted that these countries could be grouped distinctively in a given cluster based on this condition, but the variation in their classifications is also noticeable. For example, Bangladesh has been three times in Cluster 1 (the best performing cluster), but it has also been classified in Cluster 2 (deficient report of education finance's indicators) and Cluster 4 (low reporting in all dimensions). This list of 44 countries proves us that countries can exhibit complex behaviours of questionnaire response across time, but the fact that a given country has been classified at least once in Clusters 1 or 2 opens the possibility for improving their performance.

Table 42. Countries with maximum 3 times in a given cluster, excluding countries analyzed in the previous conditions (4 or more times in a given cluster)

	Number of times in (1999, 2002, 2004-2008)				
Country	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Bangladesh	**3**	1	0	2	1
Belize	**3**	2	2	0	0
British Virgin Islands	**3**	1	0	1	2
Cameroon	**3**	2	1	0	1
Greece	**3**	2	1	1	0
Guyana	**3**	2	0	1	1
Kuwait	**3**	1	2	0	1
Liechtenstein	**3**	1	0	**3**	0
Malta	**3**	**3**	0	1	0
Morocco	**3**	**3**	0	0	1
Paraguay	**3**	2	2	0	0
Swaziland	**3**	2	1	1	0
Great Britain	**3**	0	1	0	3
Belarus	2	**3**	1	0	1
China	1	**3**	0	**3**	0
Fiji	1	**3**	2	0	1
Ghana	0	**3**	**3**	0	1
Japan	0	**3**	1	0	**3**
Kiribati	0	**3**	1	2	1
Mozambique	1	**3**	**3**	0	0
Qatar	0	**3**	2	1	1
Saudi Arabia	0	**3**	0	**3**	1
Tuvalu	0	**3**	0	**3**	1
UR Tanzania	0	**3**	**3**	0	1
US America	1	**3**	1	0	2
Venezuela	2	**3**	2	0	0
Honduras	0	2	**3**	2	0
Maldives	0	2	**3**	0	2
Marshall Islands	1	0	**3**	2	1
Mauritania	1	2	**3**	0	1
Namibia	1	2	**3**	0	1
Nigeria	0	2	**3**	2	0
Samoa	1	0	**3**	**3**	0
Senegal	2	1	**3**	0	1
Sri Lanka	0	0	**3**	**3**	1
Syrian Arab Republic	0	2	**3**	0	2
Tonga	1	1	**3**	2	0
Bermuda	1	1	0	**3**	2
Bhutan	2	1	1	**3**	0
DR Congo	0	2	1	**3**	1
Lebanon	2	1	1	**3**	0
Nauru	0	2	0	**3**	2
Solomon Islands	0	1	0	**3**	**3**
Turks and Caicos Islands	2	0	1	**3**	1

CHAPTER 8. Longitudinal analysis

An important issue of interest for the parties involved in the production and analysis of internationally comparable education statistics (e.g. UNESCO, education analysts, donors, data users, governments, etc.) is the evolution of the production of statistics around the world through time. Another important issue is the link between governance and statistical capacity building (Morrison, 2005) as one of the ultimate objective of data production is to encourage the use of evidence-based policies. This information could be used for proposing new areas of work related to capacity building at the national level (e.g. priority work on tertiary education data, etc.), and building diagnostic tools for education statistical capacities.

This section studies the effects of relevant national-level indicators of governance (e.g. government effectiveness, rule of law, etc.) on the response rates of countries as measured by scores related to the five dimensions of the 45-item scale representing the structure of responses in the education database. A multilevel multinomial logistic regression model is used to take into consideration the longitudinal nature of the data.

The results show that the rate of reporting data decreases in certain dimensions (tertiary education data and detailed statistics on enrolment/teaching staff for primary and secondary education) and that the tradition of good governance has a significant effect on the production of international education statistics.

Longitudinal Analysis

A relevant feature of the production of international education statistics such as the UIS database is that it is a process that considers measures in annual bases, or in other words, it is an exercise that searches to measure the state of education in each country each year, yielding longitudinal data.

Singer and Willett (2003 : 9) mention that in order to carry an analysis of change in longitudinal data, there are certain conditions that a study must have: "three or more waves of data, an outcome whose values change systematically over time [and] a sensible metric for clocking time". As seen before, in the current study of response rates there are 209 countries (subjects) reporting data at repeated times (since 1999). The five dimensions related to the scale presented in Section 6.3 offer an efficient structure for analyzing the database response rate (or missing values) and constitute for the present study the outcomes or response variables of interest [Note: each dimension of the scale gets a score based on the sum of valid responses to the items (each valued as 1) divided by the total number of items in the dimension].

8.1 Objectives

The main objectives of this study are:

1) to characterize the changes in the scores across time for each of the five dimensions that represent the structure of the response rates; and
2) to determine if the measures of governance (as defined by the WGI project) are linked with changes on the scores.

The purpose of this analysis is to encourage the discussions about the state of international data reporting. Therefore, it is done as an exploratory endeavour rather than developed or aimed at proving strong proposition or theories.

8.2 Response variable (variable of interest) and explanatory variables

The variables of interest are the countries' score across years on each of the five dimensions of the 45–item scale related to the structure of responses in the education database. An item has a value of 1 if the data point it represents from the education database is present (observed, UIS estimation or national estimation), otherwise the value of the item is 0 (missing). Table 43 shows a brief description of each of the five dimensions. Education data (dependent variable) correspond to case A – observed values and UIS and national estimations – as defined in Section 3.1.5.

Each factor has a score each year. For example, for Factor 1 (QA) has:

$$\text{Score Factor 1}_{cj} = \frac{\sum_{k=1}^{12} (ITEM_{ckj})}{12}$$, where $ITEM_{ckj}$ corresponds to the value of the k^{th} item on the j^{th} year for the country c^{th}, and 12 is the total number of items in this dimension. In this regard, $ITEM_{kj}$ is always 0 or 1, and the range of values of a score is from 0 to 1.

Table 43. Brief description of the five dimensions from the 45-item scale

Dimension	# items	Definition of the dimension	UIS education quest.	Questionnaire subject
Factor 1	12	Detailed by grade enrolment/repeaters statistics of primary and secondary	A	Pre-primary, primary and secondary statistics
Factor 2	8	General raw data enrolment/teaching statistics of primary and secondary	A	Pre-primary, primary and secondary statistics
Factor 3	10	Net enrolment rate/gross enrolment rate/children out of school statistics of primary and secondary	A	Pre-primary, primary and secondary statistics
Factor 4	10	Educational expenditure	B	Education finance – all education levels
Factor 5	8	Tertiary education statistics	C	Tertiary education statistics

National statistical systems exist within a national context and as such, they may be affected by a multitude of factors, such as political environment, strength of public institutions, etc. It is not surprising to find that the increased demand for reliable statistics and technical assistant for statistical capacity building has been linked to the "heightened emphasis worldwide on good governance, transparency, and accountability" (Morrison, 2005).

The second objective of the present study is to determine if specific measures of governance have an impact on the countries' response rates. The measures chosen to represent governance at national levels come from the Worldwide Governance Indicators (WGI) project, developed by the Macroeconomics and Growth Team, Development Research Group, World Bank (Kaufmann *et al.*, 2010; WGI data 2011 update). The WGI project presents measures on six dimensions of governance for over 200 countries and territories for years 1996, 1998, 2000 and 2002 to 2010. This dataset is freely available.

The WGI project's work on governance is based on a concise definition that highlights three main aspects:

"[Governance is] the traditions and institutions by which authority in a country is exercised. This includes (a) the process by which governments are selected, monitored and replaced; (b)

the capacity of the government to effectively formulate and implement sound policies; and (c) the respect of citizens and the state for the institutions that govern economic and social interactions among them." (Kaufmann *et al.* 2010. p. 4)

Kaufmann *et al.* constructed two measures of governance for each of the three areas that their definition of governance encompasses. These measures are:

"(a) The process by which governments are selected, monitored, and replaced:

1. Voice and Accountability (VA) – capturing perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media.

2. Political Stability and Absence of Violence/Terrorism (PV) – capturing perceptions of the likelihood that the government will be destabilized or overthrown by unconstitutional or violent means, including politically motivated violence and terrorism.

(b) The capacity of the government to effectively formulate and implement sound policies:

3. Government Effectiveness (GE) – capturing perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies.

4. Regulatory Quality (RQ) – capturing perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development.

(c) The respect of citizens and the state for the institutions that govern economic and social interactions among them:

5. Rule of Law (RL) – capturing perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence.

6. Control of Corruption (CC) – capturing perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests." (Kaufmann *et al*, 2010, p. 4.)

Governance indicators are the result of the standardization and summarization of many perception-based governance data sources across the world. The governance indicators are reported in the standardized form (mean zero and standard deviation of one across countries per

year), and their values range approximately from -2.5 (the worst case) to 2.5 (the best case). Kaufmann *et al*, (2010) comment that although it is not possible to obtain a trend in the global averages of these indicators (values are centered at zero each year), there is very little evidence of changes in world average trends of the data sources and that fixing the global average of the governance indicators to zero is not unreasonable.

A second set of explanatory variables, used in the present study mainly as control variables, includes four statistics from the World Development Indicators (WDI) dataset. They are freely available in the World Bank website (World Bank Data Catalogue). These indicators are: Gross Domestic Product (GDP) per person, Labor participation rate, Total population and Urban population (% from the total) (for more details, see Table 44). Basically, these indicators are related to the population and the development of a country. GDP per person and, in a certain measure, labor participation rate are used as a proxy of the distribution of wealth in a country. Although a better indicator of the finance component of education statistical operations is the real budget (or expenditure) assigned by, for example, the Ministry of Education to education data collection, it is important to note that the information on operational expenditure does not seem widely available. The indicators related to population follow a simpler logic: education statistics are usually based on counting number of students, teachers, schools, which in a certain measure depend on the dynamics imposed by the education demands of the country's population.

Table 44. Details of explanatory variables related to population and development

Indicator Code (World Bank)	Indicator name	Source Note
SL.GDP.PCAP.EM.KD	GDP per person employed (constant 1990 PPP $)	GDP per person employed is gross domestic product (GDP) divided by total employment in the economy. Purchasing power parity (PPP) GDP is GDP converted to 1990 constant international dollars using PPP rates. An international dollar has the same purchasing power over GDP that a U.S. dollar has in the United States. Source: World Bank Data Catalogue
SL.TLF.CACT.ZS	Labor participation rate, total (% of total population ages 15+)	Labor force participation rate is the proportion of the population ages 15 and older that is economically active: all people who supply labor for the production of goods and services during a specified period. Source: World Bank Data Catalogue
SP.POP.TOTL	Population, total	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship--except for refugees not permanently settled in the country of asylum, who are generally considered part of the population of their country of origin. The values shown are midyear estimates. Source: World Bank Data Catalogue
SP.URB.TOTL.IN.ZS	Urban population (% of total)	Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects. Source: World Bank Data Catalogue

8.2 Statistical model

Our dependent variables, Countries' scores on the five dimensions, are based on data produced annually. Therefore, it can be expected that, for each country, these scores (observations) are correlated rather than independent, even after controlling for the explanatory variables (Hedecker, 2004). Moreover, the sphericity assumption is unlikely to hold in this longitudinal data: scores closer in time (e.g. from one year to another) are more highly correlated than scores further apart (e.g. from 1999 and 2009) (Der and Everitt, 2002). Consequently, traditional

models like linear regression, control charts, ANOVA, etc. are not the most suitable models for the study of longitudinal data.

Based on the nature of the data, a suitable modeling approach must consider parameters that take into account the structure of the repeated measures as well as parameters that link the explanatory variables with the repeated response variables (in this case, with the scores on each dimension). Singer and Willett (2003) mention that an appropriate model for studying changes must consider research questions related to *within-person change* and *between-persons differences in change.* This is referred by these authors as a two-level model: level 1 describing the manner individuals change in time, and level 2 describing the difference in change across individuals (which can include the explanatory variables). This modelling approach is also known as a *multilevel, hierarchical, random-effects or mixed models* (Hedecker, 2004).

Der and Everitt (2002 : 235) offer a short description of the use of linear mixed models for longitudinal data (or multilevel model for repeated measures):

"Linear mixed models introduce the needed correlations by formalizing the idea that an individual's pattern of responses is likely to depend on many characteristics of that individual, including some that are unobserved. These unobserved variables are then included in the model as random variables, that is *random effects*. The essential feature of such models is that correlation among the repeated measurements on the same individuals arises from shared unobserved variables, but conditional on the values of the random effects, the repeated measurements are assumed to be independent, the so-called *local independence assumption*...So in linear mixed models the mean response is modelled as a combination of population characteristics that are assumed to be shared by all individuals (the fixed effects) and subject-specific effects that are unique to a particular individual (the random effects)."

The differentiation between random and fixed effects is essential for the correct specification of the model – including the correct coding in the respective statistical software – as well as for the interpretation of results.

A simple approach to model linear growth is the random intercept and slope model [a simpler model is the random intercept model (where slopes are equal across individuals, in this case, countries), but based on the previous analyses it can be assumed that countries display different trends on the scores due to individual unobservable characteristics or random effects].

As presented by Singer (1998), in the case of a continuous response variable (Y_{ij}), for the i^{th} country on the j^{th} year, we can write the linear mixed model in two levels:

$$Y_{ij} = \pi_{0j} + \pi_{1j} TIME_{ij} + r_{ij}, \quad \text{where } r_{ij} \sim N(0, \sigma^2) \text{ (level 1 or within-subjects)}$$

$$\begin{matrix} \pi_{0j} = \beta_{00} + u_{0j}, \\ \pi_{1j} = \beta_{10} + u_{1j}, \end{matrix} \quad \text{where } \begin{pmatrix} u_{0j} \\ u_{ij} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right] \text{ (level 2 or between-subjects).}$$

The intercept π_{0j} and the slope π_{1j} allow each country to have different intercepts and slopes (time trends) by considering the addition of unique individual contributions (random effects) u_{0j} and u_{1j} to the population intercept and trend determined by β_{00} and β_{10} respectively. As seen in the between-subject equation, the distribution of the population of random effects (individual-specific) u_{0j} and u_{1j} is considered to be bivariate normal. Also, as in the case of linear regression, r_{ij} is the error term distributed normally with mean 0 and variance σ^2 with conditional independence on the values of the random effects.

Combining level 1 and level 2 equations allows separating fixed effects (first bracket) from random effects (second bracket).

$$Y_{ij} = [\beta_{00} + \beta_{10} TIME_{ij}] + [u_{0j} + u_{1j} TIME_{ij}] + r_{ij} \quad \text{(combined form)}$$

The multilevel model also allows for the inclusion of covariates (COVAR) at level 2:

$$\begin{matrix} \pi_{0j} = \beta_{00} + \beta_{01} COVAR_j + u_{0j}, \\ \pi_{1j} = \beta_{10} + \beta_{11} COVAR_j + u_{1j}, \end{matrix} \quad \text{where } \begin{pmatrix} u_{0j} \\ u_{ij} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right],$$

and the combined form is (no interaction included):

$$Y_{ij} = [\beta_{00} + \beta_{10} TIME_{ij} + \beta_{01} COVAR_j] + [u_{0j} + u_{1j} TIME_{ij}] + r_{ij}.$$

In the present study, a multilevel proportional odds model will be used. For the case of a ordered increasing responses coded with C categories = 0, 1, 2, ...C, the cumulative probabilities for the C categories of the ordinal dependent variable Y are defined as $P_{ijc} = \Pr(Y_{ij} < c) = \sum_{k=1}^{c} P_{ijk}$. The cumulative logits of this model is:

$$Log\left[\frac{P_{ijc}}{1 - P_{ijc}}\right] = \pi_{0j} + \pi_{1j} TIME_j \quad (c = 1, \ldots, C\text{-}1).$$

In this case, the level 2 can be specified as in the case of linear dependent variables.

Models for analysis

The scores' histograms show a large number of zeros and ones (a great proportion of countries were responding all items in a dimension, or not responding at all). In addition, the results of linear regressions following a logit transformation of the scores (as dependent variable) on the explanatory variables showed serious problems in the assumptions of normality of the residuals. This prevented the scores from being treated as a continuous variable in a linear regression model. Another approach is to see the scores as categories of response, and then to apply multinomial logistic regression. To fit this model, the scores of each dimension were transformed into categorical variables using the following rule: scores from 0 to 0.2 as category 0, from 0.2 to 0.5 as category 1, from 0.5 to 0.8 as category 2 and from 0.8 to 1 as category 3.

To assess the decrease, increase or lack of change of the transformed scores and to explore the significance of governance, a multilevel multinomial logistic regression model was built for each of the five factor's score. The multinomial logistic regression takes advantage of the fact that the transformed score (4 categories) can be considered as an ordinal scale. The longitudinal aspect of the data collection is captured through random effects in a two-level hierarchical model.

As recommended by Singer (1998), all explanatory variables, including control variables, were centered at the grand mean. The measures of GDP per capital and Total Population were previously transformed to a logarithmic scale. As a consequence, the estimations of the score's slopes represent the average trend of the world (excluding the countries from which governance indicators are not available, which in general are small territories or islands). Scores (variables of interest) are available since 1999 and governance indicators are available mainly since 2002, but years scores in 2009 and 2010 do not seem to be complete in the sense that their collection are still in process. Therefore, the timeframe for the analysis is set from 2002 to 2008. The variable Year starts at zero, which represents 2002.

Because one of the main objectives is to assess the significance of governance, no interactions are included in the analyses. All models include random intercepts and slopes as part of the linear country growth model. These random effects are independent from the country-level covariates. Regarding the independent variables, four models will be analyzed:

- Full Model: includes all World Government Indicators (WGI) and the control variables.

- Model A: includes indicators related to the selection, monitoring and replacement of a government (Voice & Accountability and Political Stability and Absence of Violence/Terrorism) and the control variables.
- Model B: includes indicators related to government's capacity for efficient policy's implementation (Government Effectiveness and Regulatory Quality) and the control variables.
- Model C: includes indicators related to citizens' respect to economic and social institutions and the control variables.
- Model D: includes only random intercepts and slopes.

Models A, B and C are each related to one of the three aspects encompassed by definition of governance from the WGI project.

SAS' procedure GLIMMIX was used for all models. The syntax to fit the full model for scores on the factor 5 as dependent variable is:

```
proc glimmix data=multinomial   method=laplace;
class obs;
nloptions maxiter=100;
model ScoreF5_trans(desc)   = year VoiceAcct PolStabNoViol GovtEffect
RglQlty Rulelaw CntrlCorr LGDPcap Lgpoptot Labrt Urbpop  /
dist =multinomial link = cumlogit ddfm=bw solution;
random intercept year/ subject=obs type = un;
  covtest "need random intercept?"  0 0 0;
  covtest "need random slope?" . 0 0;
run;
```

The "method = Laplace" statement indicates that the procedure GLIMMIX uses the Laplace approximation of the marginal likelihood[6], allowing for fit statistics based on (possibly restricted) log likelihood, instead of pseudo- and quasi-likelihood estimations. The latter would not be comparable across models.

It is not possible to model residual effects such as autocorrelation with this model, but instead it is expected that the random effects from intercept and slopes capture that structure. The option Dist (distribution) is specified as multinomial, and the option link (link function) is specified as cumlogit (cumulative logit) to fit a multinomial logistic regression[7]. The underlying assumption

---

[6] See GLIMMIX procedure: Fit Statistics (SAS Institute Inc, 2009 : 2276) and Maximum Likelihood with Laplace Approximation (SAS Institute Inc., 2009 : 2102).

[7] GLIMMIX procedure uses the logits of cumulative probabilities for ordinal data (SAS Institute Inc., 2009 : 2660).

of this model is referred as "proportional odds" or "equal slopes" and implies that the effect of the covariates are the same across categories of the dependent variable.

In the RANDOM statement, intercept and year are declared to specify random effects in the intercepts and slopes for time (variable year). The option Type is used to model the structure of variance covariance matrix of random effects and it is declared as UN (unstructured), which allows for separate variance components.

The COVTEST statement[8] displays the statistical inference for random intercept and random slopes.

8.4 Multinomial longitudinal analysis: results and comments

The estimations of the fixed effects for each factor are shown in Tables 45 to 49, representing the results from the multinomial logistic regression of scores from Factors 1 to 5 respectively. Let's recall that each transformed score has four categories, with category 0 as base, indicating the lowest scores possible (< 0.20), while category 4 indicates the highest score possible (> 0.80).

In the Full Model, which includes all six governance indicators, the VIF of governance indicators are rather large, ranging from 3 to 20, and denoting a possible problem of multi-collinearity. The VIF are less than 5 for Model A, and 11 or less for Models B and C. As a consequence, we draw conclusion from Models A, B and C, while the Full Model serves just as a reference.

The first objective, to characterize change across time, can be responded by examining the estimation of fixed effects of the variable Year, which represent the average logit change (slope) of scores due to time, conditional to random effects. The estimations of slopes are only significant for Factor 1 and Factor 5, which are related to detailed enrolment statistics for primary/secondary/post-secondary, and with tertiary education, respectively.

For Factor 1, the odds of being at a higher category relative to being in a given category or below is 0.78 [exp(-0.238)], in other words, scores for Factor 1 are decreasing in time. For Factor 5, the same odds are 0.80, indicating that the scores for Factor 5 are also decreasing in time. For Factor 2, 3 and 4, there is no evidence that the odds are changing through time. The estimates of slopes are identical in signs and similar in values when comparing the Full Model to Models A to D.

---

[8] See GLIMMIX procedure: COVTEST statement (SAS Institute Inc. 2009 : 2127).

The decreasing production of detailed statistics in primary/secondary and tertiary education statistics may be related to specific aspects of their nature. Detailed statistics (e.g. distribution of enrolment by grade and by age, etc.) may be more difficult to produce and may imply more specialized training for their use than gross statistics (e.g. total enrolment in secondary). As for tertiary education statistics, these are usually collected by a ministry or national authorities (e.g. ministry of higher education, national council of universities, etc.) that are independent from those in charge of collection of primary/secondary statistics (e.g. Ministry of Education, etc.). Therefore, the expertise in data collection or the investment in statistical capacity building may not be efficiently transferred to the collection of tertiary data, in turn decreasing the probabilities of response. It is worth to mention that, as with tertiary education statistics, the collection of data on education finance and expenditure has its own challenges, for example, diverse sources of information, specialized knowledge of the budgetary system, expertise in the alignment of cost and education programs, etc. Nevertheless, the scores on the factor related to education finance data (Factor 4) seem to be stable in time, a stability that may be due to important investments on capacity building at national or international scale[9]. These results also indicate that priority work must be carried in order to correct the decreasing trends in data production for Factors 1 and 5. Some suggestions are: identification of countries that are displaying problems in the collection of these data, interviews with national authorities to establish the state of production of these statistics, comparisons in production and use of these decreasing statistics versus other more stable statistics, comparisons of statistical capacities between countries, transfers of expertise from high performing countries to countries in the process of developing statistical capabilities, etc.

The second objective, to determine if governance is linked to changes in the scores, can be responded by examining the estimations of the fixed effects of the governance indicators, which are also related to the change in odds in the response variable, conditional to random effects. We can see that improvements in governance, as defined by an increase in the values of governance indicators, have significant positive effect on the scores of all factors, except for Factor 2. In general, as a unit of a governance indicator increases, the odds of being at a higher category relative to being in a given category or below increase, varying between 1.5 (for the case of Political Stability in Factor 4, Model B) and 6.5 (for the case of Rule of Law in Factor 3, Model C) with p-value < 0.05.

---

[9] An example of international capacity building is the report "Financing Education in Sub-Saharan Africa" (UIS-UNESCO, 2011), which included field work destined to build and sustain statistical capacity for reporting education finance data at the regional level.

Regarding the three aspects of governance proposed by the WGI project, Models A, B and C have governance indicators that are significant, but their effects depend on the studied factors. For example, for Factor 1, the effects of Voice and Accountability in Model A, and Rule of Law in Model C, are significant, but none of the governance indicators related to Model B (Government Effectiveness and Regulatory Quality) are significant. Nevertheless, this is not the case for scores in Factor 4, where Models A, B and C have one significant indicator each at p-value < 0.05. Moreover, at p-value < 0.10, Model A, B and C have one significant governance indicator for Factor 3.

The positive effect of good governance on the education data production is not unexpected. Governance is intrinsically related to the strength of national institutions, which in turn can affect the different outputs that they deliver, including production of national statistics and international reports. Data on education from primary to tertiary levels, including education expenditures, are essential for efficient policy making. It would be therefore relevant to find out if improvements on governance are a consequence of premeditated actions by the part of politicians and civil society and if there are elements in these actions that affect directly the production of education statistics (e.g., training in the use of statistics for policy making, the compromise to fill out international reports, etc.). Investments on actions that impact both governance and statistical capacities can improve the production of statistics and the demand for them by users, creating a sustainable process where the benefits of evidence-based policies are higher than the cost of collecting, processing and disseminating relevant data.

Regarding the effects of the control variables, it is interesting to note that the logarithm GDP per capita has a significant effect (negative log odds) only for Factor 1, while the logarithm of total population has significant effect for Factors 4 and 5, and the proportion of urban population has significant effect for Factors 1 and 3. These control variables are not subject to easy manipulations; nevertheless, it is important to note that good results in data collection may not be correlated to levels of the GDP per capita, in other words, improvement in the report of education statistics may not depend on the level of development of a country, but may depend on other important factors that are easier to change, such as political will, governance, user's demands, operational budget, etc.

We can note that the addition of control variables and governance indicators (Models A, B and C) decreases the value of AIC, AICC and BIC in comparison to those values from the model with only random intercept and slopes (Model D). Also, all the test of the random intercept and slopes reject the null hypothesis that they are zero, indicating that the choice of modeling a

multilevel model is appropriate to the characteristics of the dataset. In addition, a model whose dependent variable was the transformation of the scores in each factor into logits was also fit using SAS' proc MIXED. The results showed that the log odds estimations for the trends are very similar to the corresponding parameters in proc MIXED in magnitude, sign and statistical significance. The log odds estimations of the governance indicators also display also some similarities in magnitude.

Table 45. Multilevel multinomial (proportional odds) logistic regression for Factor 1

	Full Model	Model A	Model B	Model C	Model D
Trend					
Year	-0.238 *	-0.247 *	-0.242 *	-0.239 *	-0.247 **
	(0.106)	(0.103)	(0.104)	(0.103)	(0.074)
Governance Indicators					
Voice & Accountability	0.844 *	0.948 **			
	(0.416)	(0.349)			
Political Stability	-0.327	-0.196			
	(0.367)	(0.34)			
Government Effectiveness	0.796		0.863		
	(0.716)		(0.584)		
Regulatory Quality	-0.474		-0.110		
	(0.609)		(0.573)		
Rule of Law	1.085			1.507 *	
	(0.717)			(0.595)	
Control of Corruption	-1.044 +			-0.743	
	(0.587)			(0.528)	
Control variables					
Lg GDP per cap	-0.788 *	-0.631 *	-0.725 *	-0.773 *	
	(0.366)	(0.321)	(0.344)	(0.343)	
Lg Population	-0.131	-0.090	-0.116	-0.071	
	(0.176)	(0.17)	(0.157)	(0.158)	
Labor participation rate	0.021	0.017	0.010	0.013	
	(0.031)	(0.031)	(0.03)	(0.03)	
Urban population (%)	0.046 *	0.040 *	0.038 *	0.043 *	
	(0.019)	(0.018)	(0.018)	(0.018)	
Fit Statistics					
-2 Log Likelihood	1941.75	1947.06	1953.33	1950.21	2496.53
AIC	1975.75	1973.06	1979.33	1976.21	2510.53
AICC	1976.27	1973.37	1979.64	1976.52	2510.61
BIC	2029.06	2013.82	2020.1	2016.97	2533.93

Note: GDP = Gross Domestic Product. Lg = Logarithmic transformation.

Urban population (%) = Proportion of urban population from the total population.

+ *p-value* < 0.10; * *p-value* < 0.05; ** *p-value* < 0.001

Table 46. Multilevel multinomial (proportional odds) logistic regression for Factor 2

	Full Model	Model A	Model B	Model C	Model D
Trend					
Year	0.074	0.056	0.079	0.022	0.039
	(0.071)	(0.067)	(0.069)	(0.069)	(0.054)
Governance Indicators					
Voice & Accountability	0.144	0.114			
	(0.366)	(0.31)			
Political Stability	0.546 +	0.358			
	(0.319)	(0.297)			
Government Effectiveness	1.560 *		0.666		
	(0.652)		(0.53)		
Regulatory Quality	0.422		0.004		
	(0.547)		(0.511)		
Rule of Law	-1.865 **			-0.536	
	(0.669)			(0.535)	
Control of Corruption	-0.181			0.427	
	(0.511)			(0.474)	
Control variables					
Lg GDP per cap	0.022	0.116	-0.060	0.389	
	(0.33)	(0.286)	(0.321)	(0.32)	
Lg Population	-0.123	-0.035	-0.129	-0.106	
	(0.156)	(0.152)	(0.144)	(0.145)	
Labor participation rate	0.021	0.020	0.019	0.024	
	(0.027)	(0.027)	(0.027)	(0.028)	
Urban population (%)	-0.001	0.000	0.000	-0.004	
	(0.017)	(0.017)	(0.016)	(0.017)	
Fit Statistics					
-2 Log Likelihood	2224.36	2236.49	2236.84	2238.95	2892.45
AIC	2258.36	2262.49	2262.84	2264.95	2906.45
AICC	2258.89	2262.8	2263.15	2265.26	2906.53
BIC	2311.67	2303.26	2303.6	2305.72	2929.85

Note: GDP = Gross Domestic Product. Lg = Logarithmic transformation.

Urban population (%) = Proportion of urban population from the total population.

+ *p-value* < 0.10; * *p-value* < 0.05; ** *p-value* < 0.001

Table 47. Multilevel multinomial (proportional odds) logistic regression for Factor 3

	Full Model	Model A	Model B	Model C	Model D
Trend					
Year	-0.102	-0.145	-0.085	-0.097	-0.100
	(0.101)	(0.1)	(0.1)	(0.101)	(0.085)
Governance Indicators					
Voice & Accountability	1.269 **	1.674 **			
	(0.428)	(0.367)			
Political Stability	-0.459	-0.238			
	(0.364)	(0.345)			
Government Effectiveness	1.672 *		1.864 **		
	(0.736)		(0.617)		
Regulatory Quality	-0.380		-0.001		
	(0.618)		(0.592)		
Rule of Law	0.111			1.220 +	
	(0.734)			(0.627)	
Control of Corruption	-0.236			0.388	
	(0.61)			(0.562)	
Control variables					
Lg GDP per cap	-0.535	-0.127	-0.562	-0.362	
	(0.385)	(0.346)	(0.391)	(0.389)	
Lg Population	0.116	0.229	0.125	0.238	
	(0.193)	(0.19)	(0.182)	(0.185)	
Labor participation rate	0.000	0.003	-0.007	-0.008	
	(0.033)	(0.033)	(0.033)	(0.034)	
Urban population (%)	0.042 *	0.039 +	0.038 +	0.036 +	
	(0.021)	(0.021)	(0.021)	(0.022)	
Fit Statistics					
-2 Log Likelihood	1731.46	1738.83	1745.15	1751.01	2332.14
AIC	1765.46	1764.83	1771.15	1777.01	2346.14
AICC	1765.98	1765.14	1771.46	1777.32	2346.22
BIC	1818.77	1805.59	1811.91	1817.78	2369.54

Note: GDP = Gross Domestic Product. Lg = Logarithmic transformation.

Urban population (%) = Proportion of urban population from the total population.

+ *p-value* < 0.10; * *p-value* < 0.05; ** *p-value* < 0.001

Table 48. Multilevel multinomial (proportional odds) logistic regression for Factor 4

	Full Model	Model A	Model B	Model C	Model D
Trend					
Year	0.040	-0.002	0.025	0.018	-0.033
	(0.057)	(0.055)	(0.057)	(0.056)	(0.046)
Governance Indicators					
Voice & Accountability	0.804 **	1.225 **			
	(0.298)	(0.255)			
Political Stability	0.171	0.414 +			
	(0.267)	(0.249)			
Government Effectiveness	0.259		1.062 *		
	(0.582)		(0.466)		
Regulatory Quality	0.190		0.676		
	(0.476)		(0.447)		
Rule of Law	0.297			1.021 *	
	(0.567)			(0.462)	
Control of Corruption	0.390			0.688 +	
	(0.438)			(0.413)	
Control variables					
Lg GDP per cap	-0.462 +	-0.141	-0.397	-0.338	
	(0.26)	(0.227)	(0.257)	(0.251)	
Lg Population	0.358 **	0.413 **	0.245 *	0.350 **	
	(0.126)	(0.124)	(0.113)	(0.117)	
Labor participation rate	-0.002	0.005	0.000	-0.002	
	(0.022)	(0.022)	(0.021)	(0.022)	
Urban population (%)	0.021	0.021	0.019	0.018	
	(0.013)	(0.013)	(0.013)	(0.013)	
Fit Statistics					
-2 Log Likelihood	2307.06	2314.25	2324.07	2321.25	2793.76
AIC	2341.06	2340.25	2350.07	2347.25	2807.76
AICC	2341.59	2340.56	2350.38	2347.56	2807.83
BIC	2394.37	2381.02	2390.83	2388.02	2831.15

Note: GDP = Gross Domestic Product. Lg = Logarithmic transformation.

Urban population (%) = Proportion of urban population from the total population.

+ *p-value* < 0.10; * *p-value* < 0.05; ** *p-value* < 0.001

Table 49. Multilevel multinomial (proportional odds) logistic regression for Factor 5

	Full Model	Model A	Model B	Model C	Model D
Trend					
Year	-0.203 *	-0.244 **	-0.204 *	-0.231 **	-0.168 **
	(0.085)	(0.082)	(0.083)	(0.083)	(0.065)
Governance Indicators					
Voice & Accountability	0.005	0.432			
	(0.421)	(0.372)			
Political Stability	0.103	0.265			
	(0.353)	(0.332)			
Government Effectiveness	0.503		0.195		
	(0.723)		(0.603)		
Regulatory Quality	1.182 *		1.090 +		
	(0.594)		(0.574)		
Rule of Law	0.614			1.333 *	
	(0.705)			(0.595)	
Control of Corruption	-1.101 +			-0.543	
	(0.574)			(0.518)	
Control variables					
Lg GDP per cap	0.138	0.561	0.182	0.416	
	(0.4)	(0.35)	(0.387)	(0.387)	
Lg Population	0.365 +	0.460 *	0.368 *	0.425 *	
	(0.191)	(0.189)	(0.178)	(0.179)	
Labor participation rate	-0.029	-0.020	-0.033	-0.022	
	(0.033)	(0.033)	(0.033)	(0.034)	
Urban population (%)	0.006	0.001	0.001	0.004	
	(0.02)	(0.02)	(0.02)	(0.02)	
Fit Statistics					
-2 Log Likelihood	2167.6	2177.57	2174.3	2177.56	2673.52
AIC	2201.6	2203.57	2200.3	2203.56	2687.52
AICC	2202.13	2203.88	2200.61	2203.87	2687.6
BIC	2254.91	2244.33	2241.07	2244.32	2710.92

Note: GDP = Gross Domestic Product. Lg = Logarithmic transformation.

Urban population (%) = Proportion of urban population from the total population.

+ *p-value* < 0.10; * *p-value* < 0.05; ** *p-value* < 0.001

CHAPTER 9. Conclusion

The central idea of the present research is to explore the situation of UIS education database from the point of view of data production (response rates and missing values) and to look into some causes and effects.

The importance of the UIS education database in the context of international development is discussed in the literature review (Chapter 2). It was noted that the negative effects of missing values can be felt in cost increases of data processing and validation and in possible loss of inferential power in the monitoring analysis carried by UIS. It was also noted that the problems related to missing data can be traced to problems in the reliability of statistical production capacities at the national level.

To fully grasp the phenomena of completeness (response rate/missing values) in the data base, many types of analysis were done, each of them focusing on a different aspect.

The education statistics used in this study correspond to academic years 1999 to 2008. Data for academic years 2009 and 2010 were still under collection during the present study. In addition, there is some evidence that the 2008 global economic crisis affected the data production at country level in 2009 (in my work as statistical assistant, certain countries' respondents expressed this opinion); in this regard, it is recommendable to study the production of education data for 2009 and 2010 in an independent manner. Indeed, the study of the impact of the economic crisis on the production of education statistics and, in general, the national statistical system could be useful for understanding the different factors that may influence in the robustness or reliability of the production of statistics worldwide.

From the descriptive analysis (Chapter 3), it was noted that, considering the ensemble of the education database (over 500 data and indicators), the quantity of data available for dissemination has a "reporting" peak in the academic years 2004 and 2005 (around 60% of completeness of the database), while for academic year 2008 the completeness is slightly less (57.7%), indicating a possible declining trend. At the same time, the production of UIS estimations seems to be decreasing: it went from over 9000 estimated data points for academic year 2003 to around 2300 for academic year 2008. The causes for these decreases in the available data and the UIS estimations were not identified, but it is worth noting that the UIS international data collection is indeed a very challenging process that involved the participation of many parties at the national and international levels and that tracing any chronic cause of bad

quality needs careful revision of internal processes of data production as well as intense statistical capacity building work in member states.

Also, it was noted that, on average, between 60% and 70% of countries have response rates, for the ensemble of the database, that vary from 50% to 90%. In other words, there is still work to do with approximated 40% of countries that cannot produce/submit enough data to at least complete 50% of their expected education statistics and indicators (as disseminated in the UIS database). Variables with the best and worst response rates were identified. Among the worst variables in terms of completeness – starting at 1% up to 17% of response rate – we find statistics/indicators related to participation in tertiary education (e.g. outbound mobile students, outbound mobility ratio, and gross outbound enrolment ratio), completion of primary (e.g. expected gross primary graduation rate, etc.) and teachers (e.g. percentage of trained teachers, etc.). This contrasts with variables like enrolment in primary and gross enrolment ratio for primary, which have the highest response rate (over 80% of response rate). It has been determined that multiple factors could lead to low response rates for a variable at a country level: cost-benefits issues for national statisticians, lack of expertise on production and analysis of certain variables, lack of resources for data collection, etc. Further studies on this subject are recommended.

The Statistical Capacity Indicator (SCI) scores provide an assessment at the national level of the statistical capacity of 145 countries. The comparison of the level of data completeness of these countries (country response rates) and their respective SCI scores (Chapter 4) allows us to determine which countries could be underperforming in their production of international education data with respect to their national statistical capacities (e.g. Egypt, Moldova, Albania, Slovakia, etc.). These countries constitute a good opportunity for education statistical capacity improvements, as it can be presumed that required political will and technical capabilities are already in place at the national level.

Through the analysis of trajectories of responses by subgroups of variables (linear regression and control charts; Chapter 5), we could recognize different behaviours of the response rate of variable subgroups: few groups seem to be increasing (e.g. repeaters in primary and secondary, etc.), some others seem to be decreasing (e.g. enrolment in primary, enrolment in tertiary, gross enrolment ratios, school life expectancy, teaching staff by ISCED, etc.), while the majority of variables seem to have stable response rates. Although these traditional analysis tools allow for a quick glance at the situation of the response rates, they are based on the independence of observations across time, an assumption that may not hold for the UIS data collection. A

longitudinal study of data collection, which takes into account correlation of observations, is presented in Chapter 8.

Given the great quantity of variables to analyze, many of them highly correlated, the need for the reduction of data was evident. In this regard, factor analysis and cluster analysis (Chapters 6 and 7) sought to understand the underlying structure of the response rates and to suggest a classification of countries around the world based on a proposed structure of responses. Through factor analysis, it was concluded that response patterns in the education database could be represented by 5 dimensions (slightly correlated): three related to the UIS questionnaire A (statistics on enrolment/teaching staff for pre-primary, primary, secondary and post-secondary non-tertiary), one related to UIS questionnaire B (finance) and the last one related to UIS questionnaire C (tertiary education) (Chapter 6). Moreover, this solution proved stable across time. Following the need for data reduction, a 5-dimension scale based on 45 items is proposed in order to capture and efficiently manage the variability of responses in the education database.

This view of the production of international education statistics – the study of the underlying structure – proves itself very valuable and necessary. The study of the average response rate for the whole database was certainly hiding certain patterns of responses.

Chapter 7 presents a 5-cluster classification of countries around the world based on their capacity to respond to proposed scale. The first cluster includes countries that perform well in all dimensions, in other words, countries with a developed education statistical capacity. The second cluster includes the countries that perform well in all response dimensions, except for the dimension related to education finance data. Countries in the third cluster perform well only in the first three factors (related to questionnaire A), but have problems reporting finance education and tertiary education data. The fourth cluster includes countries that do not perform well in any dimension, in other words, countries with chronic problems in their education statistical capacity. The fifth cluster includes countries that do not fit well in the previous clusters each year (in average 10% of the total). This 5-cluster classification, which is robust in time, could be the bases for future diagnostic tools and analyses, as countries in each cluster may present the same challenges regarding the construction of their statistical capacity. Moreover, after recognizing which countries have a developed education statistical capacity, the UIS could integrate them into helping in the building of education statistical capacity of neighbouring countries that may lack a stable system of education statistics, as these high performing countries may have the needed expertise or may have resolved similar problems than neighbour countries that do not perform with the same efficiency.

A relevant matter for education analysts, donors, data users and governments is the assessment of the evolution in time of the production of education statistics around the world. Chapter 8 presents the results of multinomial logistic regressions on the scores (response rates) of each of the five factors previously proposed. We can conclude that the average reports/production of detailed statistics on enrolment/teaching staff related to primary and secondary education (Factor 1) as well as the reports of tertiary education statistics (Factor 5) are decreasing in time. It can be noted that variables that belong to the subgroups of decreasing response rates, as described by the control chart analysis, also belong to the factors that show negative trends (e.g. gross enrolment ratio, enrolment in tertiary, school life expectancy from primary to tertiary, etc.).

These decreasing trends could become relevant problems in the long term; therefore, preventive and corrective actions – related to them and related to the variables with very low response rates – are necessary. Further studies at the field level are required in order to understand the problems that countries face when reporting these data. In addition, country response diagnosis can be highly improved if the classification of countries from Chapter 7 is taken into account, as it has been shown that statistical capacities across countries are not homogenous, which implies that many strategies for statistical building may be needed. Data validation, estimation and field work in general may become a priority when dealing with decreasing data. It may be necessary to find out cost-efficient procedures for data processing as well as efficient strategies to encourage countries to report complete data to the UIS.

The results regarding the decreasing trends of scores from Factors 1 and 5 take into account the structure of the response matrix as well as the correlation between observations from the same country; in this regard, these conclusions enrich and expand the preliminary results from previous chapters regarding subgroup response rates.

The longitudinal study of factor's scores also points out that increases in governance, as measure by the World Governance Indicators, has a relation with increase in reporting rates of the UIS education questionnaire. This result illustrates the positive relationship between statistics and governance, which is one of the ultimate goals of statistical production (encouraging evidence-based policies).

In the literature review, the importance of information technologies for data validation was highlighted. The analysis tools used in the present research can be incorporated into automatic templates and reports of missing data/completeness at a low cost. The automatization of

templates and reports will certainly improve the diagnostic work of analysts in charge of data processing and field work.

ANNEX

Annex 1. List of selected variables (45 items)

Code	Parent	Subgroup	Concept	World Bank education statistics	UIS Country profile	Extra	Selected 45 variables
AIRFT	Entry	Intake to primary	Gross intake ratio. Primary. Total	X			X
GPAIR	Entry	Intake to primary	Gender parity index for gross intake ratio. Primary				X
GPNIR	Entry	Intake to primary	Gender parity index for net intake rate. Primary			X	X
NINFT	Entry	Intake to primary	Net intake rate. Primary. Total			X	X
GGFG	Completion	Completion / graduates ratios	Gender parity index for gross primary graduation rate				X
GGFT	Completion	Completion / graduates ratios	Gross primary graduation rate. Total			X	X
GIRLT	Completion	Proxy completion	Gross intake ratio to the last grade of primary. Total	X	X	X	X
GPGIL	Completion	Proxy completion	Gender parity index for gross intake ratio to the last grade of primary			X	X
SR5FF	Progression	Survival	Survival rate to grade 5. Female				X
SR5FT	Progression	Survival	Survival rate to grade 5. Total		X		X
PRFF	Progression	Percentage of repeaters	Percentage of repeaters in primary. All grades. Female				X
PRFT	Progression	Percentage of repeaters	Percentage of repeaters in primary. All grades. Total	X	X		X
GPTR	Progression	Transition	Gender parity index for transition rate, primary to secondary, general programmes				X
TRANT	Progression	Transition	Transition from ISCED 1 to ISCED 2, general programmes (%). Total	X	X	X	X
TVTSP	Participation	Programme orientation	Technical/vocational enrolment in ISCED 2 and 3 as % of total enrolment in ISCED 2 and 3	X			X
PTRF	Teacher	Pupil-teacher ratio	Pupil-teacher ratio. Primary	X	X	X	X

Code	Parent	Subgroup	Concept	World Bank education statistics	UIS Country profile	Extra	Selected 45 variables
PTRS	Teacher	Pupil-teacher ratio	Pupil-teacher ratio. Secondary	X			X
GER0F	Participation	Gross enrolment ratio	Gross enrolment ratio. Pre-primary. Female		X		
GER0M	Participation	Gross enrolment ratio	Gross enrolment ratio. Pre-primary. Male		X		
GER0T	Participation	Gross enrolment ratio	Gross enrolment ratio. Pre-primary. Total	X	X		X
GERFF	Participation	Gross enrolment ratio	Gross enrolment ratio. Primary. Female		X		
GERFM	Participation	Gross enrolment ratio	Gross enrolment ratio. Primary. Male		X		
GERFT	Participation	Gross enrolment ratio	Gross enrolment ratio. Primary. Total		X		X
GERSF	Participation	Gross enrolment ratio	Gross enrolment ratio. Secondary. All programmes. Female		X		
GERSM	Participation	Gross enrolment ratio	Gross enrolment ratio. Secondary. All programmes. Male		X		
GERST	Participation	Gross enrolment ratio	Gross enrolment ratio. Secondary. All programmes. Total		X		X
GPGE0	Participation	Gross enrolment ratio	Gender parity index for gross enrolment ratio. Pre-primary				X
GPGEP	Participation	Gross enrolment ratio	Gender parity index for gross enrolment ratio. Primary	X			X
GPGES	Participation	Gross enrolment ratio	Gender parity index for gross enrolment ratio. Secondary. All programmes				X
GPNEP	Participation	Net enrolment rate	Gender parity index for net enrolment rate. Primary				X
GPNES	Participation	Net enrolment rate	Gender parity index for net enrolment rate. Secondary				X
NERFF	Participation	Net enrolment rate	Net enrolment rate. Primary. Female		X		
NERFM	Participation	Net enrolment rate	Net enrolment rate. Primary. Male		X		

Code	Parent	Subgroup	Concept	World Bank education statistics	UIS Country profile	Extra	Selected 45 variables
NERFT	Participation	Net enrolment rate	Net enrolment rate. Primary. Total	X	X	X	X
NERSF	Participation	Net enrolment rate	Net enrolment rate. Secondary. All programmes. Female		X		
NERSM	Participation	Net enrolment rate	Net enrolment rate. Secondary. All programmes. Male		X		
NERST	Participation	Net enrolment rate	Net enrolment rate. Secondary. All programmes. Total	X	X		X
SLAF	Participation	School life expectancy	School life expectancy (years). Primary to tertiary. Female				X
SLAT	Participation	School life expectancy	School life expectancy (years). Primary to tertiary. Total	X	X		X
ROFF	Participation	Out-of-school children	Rate of primary school age children out of school. Female				X
ROFT	Participation	Out-of-school children	Rate of primary school age children out of school. Total	X		X	X
PRSF	Progression	Percentage of repeaters	Percentage of repeaters in secondary. All grades. Female				X
PRST	Progression	Percentage of repeaters	Percentage of repeaters in secondary. All grades. Total	X			X
EC0TO	Expenditure	Public current expenditure	Percentage distribution of public current expenditure on education by level. Pre-primary		X		X
EC1TO	Expenditure	Public current expenditure	Percentage distribution of public current expenditure on education by level. Primary		X	X	X
ECNTO	Expenditure	Public current expenditure	Percentage distribution of public current expenditure on education not allocated by level		X		X
ECSTO	Expenditure	Public current expenditure	Percentage distribution of public current expenditure on education by level. Secondary		X		X
ECTTO	Expenditure	Public current expenditure	Percentage distribution of public current expenditure on education by level. Tertiary		X		X
EEGDP	Expenditure	Percentage of GDP / GNP	Public expenditure on education as % of GDP	X	X		X
EEGE	Expenditure	Percentage of GDP / GNP	Public expenditure on education as % of total government expenditure	X	X		X

Code	Parent	Subgroup	Concept	World Bank education statistics	UIS Country profile	Extra	Selected 45 variables
GERTF	Participation	Gross enrolment ratio	Gross enrolment ratio. ISCEDED 5 and 6. Female		X		X
GERTM	Participation	Gross enrolment ratio	Gross enrolment ratio. ISCEDED 5 and 6. Male		X		X
GERTT	Participation	Gross enrolment ratio	Gross enrolment ratio. ISCEDED 5 and 6. Total	X	X		X
PEPTF	Participation	Private education	Percentage of private enrolment. Primary	X			X
TRA1T	Teacher	Trained teacher	Percentage of trained teachers. Primary. Total	X			X

BIBLIOGRAPHY

ALLISON, Paul D. (2001). *Missing Data*, Series: Quantitative Applications in the Social Science, Vol. 136, Thousand Oaks, CA., SAGE Publications, p. 104.

BATINI, Carlo and Monica SCANNAPIECO (2006). *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*, New Jersey, Springer-Verlag, 262 p.

BBSC 2011- Excel [on-line database], Bulletin Board on Statistical Capacity, Washington, D.C., World Bank [retrieved on October 21, 2011]. < http://data.worldbank.org/data-catalog/bulletin-board-on-statistical-capacity >.

BLANK, Rolf, L. (1993). « Developing a system of education indicators: Selecting, implementing, and reporting indicators », *Educational Evaluation and Policy Analysis*, Vol. 15, No. 1 (Spring), p. 65-80.

BOTTANI, Norberto (1998). « The OECD educational indicators: Purposes, limits and production processes Prospects », *Quarterly review of comparative education*, Vol. 28, No. 1, p. 61-75.

BOTTANI, Norberto and Albert TUIJNMAN (1994). « International Education Indicators: Framework, Development, and Interpretation », in *Making Education Count: Developing and Using International Indicators,* Norberto Bottani and Albert Tuijnman (ed.), Paris, Organisation for Economic Co-operation and Development, p. 368.

BRACKSTONE, Gordon (1999). « Managing data quality in a statistical agency », *Survey Methodology*, Vol. 25, No. 2 (December), p. 139-149.

BRYK, Anthony S. and Kim L. HERMANSON (1993). « Educational Indicator Systems: Observations on Their Structure, Interpretation, and Use », *Review of Research in Education*, Vol. 19, No. 1, p. 451-484.

COX, David Roxbee and Joyce SNELL (1989). *The Analysis of Binary Data*, 2nd edition, Monographs on Statistics and Applied Probability 32, London, Chapman & Hall, 236 p.

CUSSÓ, Roser and Sabrina D'AMICO (2005). « Vers une comparabilité plus normative des statistiques internationales de l'éducation : de l'éducation de masse aux compétences », in *Pouvoirs et mesure en éducation*, A. Vinokur (dir.), Cahiers de la Recherche sur l'éducation et les savoirs, Hors-série n°1, p. 21-47.

DE VRIES, Willem F. M. (2001). « Meaningful Measures: Indicators on Progress, Progress on Indicators », *International Statistical Review / Revue Internationale de Statistique*, Vol. 69, No. 2 (August), p. 313-331.

FIELD, Andy (1998). « A bluffer's guide to ... sphericity », *The British Psychological Society: Mathematical, Statistical & Computing Section Newsletter,* Vol. 6, No. 1 p. 13-22.

GRAHAM, John W. (2009). « Missing data analysis: making it work in the real world », *Annual Review of Psychology*, Vol. 60, p. 549-576.

HAIR, Joseph F., William BLACK, Barry BABIN and Rolph ANDERSON (2009). *Multivariate Data Analysis,* 7th edition, New Jersey, Prentice Hall, 816 p.

HEDECKER, Donald (2004). « An introduction to growth modeling », in *The Sage Quantitative Methodology for the Social Science*, D. Kaplan (ed.), Thousand Oaks, CA., Sage publications, p. 215-234.

HEYNEMAN Stephen P. (2003). « The history and problems in the making of education policy at the World Bank 1960-2000 », International Journal of Educational Development, Vol. 23, No. 3 (May), p. 315-337.

HOYLE, David (2009). *ISO 9000 Quality Systems Handbook - Using the Standards as a Framework for Business Improvement (6th Edition)*, Massachusetts, Elsevier Science and Technology, 802 p.

INTERNATIONAL MONETARY FUND (2002). *The Framework for Determining Statistical Capacity Building Indicators*, background paper for the Seminar on Statistical Capacity Building Indicators, April 29-30, 2002, Washington, D.C., Statistics Department, International Monetary Fund.

JOHNSTON, John (1972). *Econometric Methods, 2nd edition*, New York, McGraw-Hill, 437 p.

KARR, Alan , Ashish SANIL and David BANKS (2006). « Data quality: A statistical perspective », *Statistical Methodology*, Vol. 3, No. 2 (April), p. 137-173.

KAUFMANN, Daniel, Aart KRAAY and Massimo MASTRUZZI (2010). *The Worldwide Governance Indicators : A Summary of Methodology, Data and Analytical Issues,* World Bank Policy Research Working Paper No. 5430 (September), Washington, D.C., World Bank.

KENNETH Ross and Ilona JÜRGENS-GENEVOIS (eds.) (2006). *Cross-national studies of the quality of education: planning their design and managing their impact*, Paris, UNESCO International Institute for Education Planning, 320 p.

KRUEGER Alan B. and Mikael LINDAHL (2000). « Education for Growth: Why and For Whom? », *Journal of Economic Literature*, Vol. 39, No. 4 (December), p. 1101-1136.

LAROCQUE, Denis (2006). *Analyse multidimensionnelle - Recueil 6602F*, Montréal, HEC Montréal, 409 p.

LEWIN, Keith M. (2011). *Taking Targets to Task Revisited: How Indicators of Progress on Access to Education can Mislead*, Research Monograph No. 54 (January), Falmer, Consortium for Research on Education Access, Transition & Equity, Centre for International Education, University of Sussex, p. 35.

MACCALLUM, Robert, Keith WIDAMAN, Shaobo ZHANG and Sehee HONG (1999). « Sample size in factor analysis », *Psychological Methods*, Vol. 4, No. 1, p. 84-99.

MCEWEN, Nelly (1990). *Educational Quality Indicators: Developing Indicator Systems in Alberta*, Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990), Alberta, Planning and Policy Secretariat, Alberta Education, 21 p.

MONTGOMERY, Douglas (2001). *Introduction to Statistical Quality Control,* 4th edition, New York, Wiley, 796 p.

MORRISON, Thomas K. (editor) (2005). *Statistical Capacity Building. Case studies and lessons learned.* Washington D.C., International Monetary Fund, 53 p.

NATIONAL RESEACH COUNCIL (1995). *Worldwide Education Statistics: Enhancing UNESCO's Role,* Washington, D.C., The National Academy Press, 66 p.

OECD (2004). *OECD Handbook for Internationally Comparative Education Statistics: Concepts, Standards, Definitions and Classifications*, Paris, OECD Publishing, 271 p.

POSTLETHWAITE, T. Neville (2004). *Monitoring educational achievement*, series: Fundamentals of educational planning Vol. 81, Paris, UNESCO International Institute for Educational Planning, 139 p.

PURYEAR, Jeffrey M. (1995). « International education statistics and research: Status and problems », *International Journal of Educational Development*, Vol. 15, No. 1 (January), p. 79-91.

RADERMACHER, Walter, Richard LAUX and Antonio BAIGORRI (2009). *Building confidence in the use of administrative data for statistical purposes*, Invited paper meeting, No. 94 [on-line], 57th session of the International Statistical Institute, Durban, South Africa, (16-22 August) [retrieved on December 20, 2011].
< http://www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/1162.pdf >.

RALPH, John (1996). « Issues in the development of national education indicators in the United States », *International Journal of Educational Research*, Volume 25, No. 3, 1996, p. 231-238.

ROWE, Ken and Denise LIEVESLEY (2002). *Constructing and using educational performance indicators*, background paper to keynote address and workshops presented at the inaugural Asia-Pacific Educational Research Association (APERA) regional conference [on-line], Melbourne, April 16-19, 2002, ACER [retrieved September 2, 2011].
< http://research.acer.edu.au/cgi/viewcontent.cgi?article=1013&context=learning_processes >.

SAS INSTITUTE INC. (2005). « Create a polychoric correlation or distance matrix », *Knowledge Base, Sample & Notes, Sample 25010* [on-line], North Caroline, SAS Institute Inc [retrieved on October 15, 2011]. < http://support.sas.com/kb/25/010.html#ref >.

SAS INSTITUTE INC. (2009). *SAS/STAT® 9.2 User's Guide, Second Edition*, Cary, North Caroline, SAS Institute Inc., 7869 p.

SCHAFER, Joseph and John W. GRAHAM (2002). « Missing data: our view of the state of the art », *Psychological Methods*, Vol. 7, No. 2 (June), p. 147-177.

SINGER, Judith (1998). « Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models », *Journal of Education and Behavioral Statistics*, Vol. 24 (winter), No. 45, p. 323-355.

SINGER, Judith D. and John WILLET (2003). *Applied Longitudinal Data Analysis: Modeling change and event occurrence*, New York, Oxford University Press, 664 p.

SKILBECK, Malcolm (2006). « A Global Endeavour: Education for All », Secondary Education at the crossroads. Education in the Asia-Pacific region: Issues, Concerns and Prospect. Vol. 9, p. 97-144.

STATISTICS CANADA (2002). *Statistics Canada's Quality assurance framework*, Catalogue no. 12-586-XIE [on-line], Ottawa, Ministry of Industry [retrieved on December 20, 2011]. < http://publications.gc.ca/Collection/Statcan/12-586-XIE/12-586-XIE2002001.pdf >.

STATISTICS CANADA (2009). *Statistics Canada Quality Guidelines, Fifth Edition*, Catalogue no. 12-539-X [on-line], Ottawa, Ministry of Industry [retrieved on December 15, 2011], 89 p. < http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf >.

STEVENS, James (2002). *Applied Multivariate Statistics for the Social Science,* 4th edition, New Jersey, Lawrence Erlbaum Assoc., 699 p.

THAGA, Keoagile (2008). « Control chart for autocorrelated processes with heavy tailed distributions », *Economic Quality Control*, Vol. 23, No. 2, p. 197-206.

UEBERSAX, John S. (2007). *TetMat – software for the estimation of matrices of tetrachoric correlations (v.1.0.3)* [on-line], [retrieved on October 20, 2011]. < http://www.john-uebersax.com/bin/tetmat.zip >.

UNDATA (2011). « Terms and Conditions of use », in *about UNdata (version v0.14.6 Beta)* [on-line], NY, United Nations Statistics Division, UN [retrieved on December 15, 2011]. < http://data.un.org/Host.aspx?Content=UNdataUse >.

UNData Data sets [on-line database], NY, United Nations Statistics Division, UN [retrieved on June 4, 2011]. < http://data.un.org/Explorer.aspx?d=UNESCO >.

UN-ECONOMIC AND SOCIAL COUNCIL (1999). « Integrated and coordinated implementation and follow-up of major United Nations conferences and summits: A critical review of the development of indicators in the context of conference follow-up », Report of the Secretary-General E/1999/11 [on-line]. NY, United Nations [retrieved on September 2, 2011]. <http://www.un.org/documents/ecosoc/docs/1999/e1999-11>.

UNESCO-UIS (2000). *Basic Texts* [on-line], Paris, UNESCO [retrieved on September 2, 2011]. < http://www.uis.unesco.org/Library/Documents/Basic%20text-en.pdf >.

UNESCO-UIS (2007). *Medium-term strategy 2008-2013* [on-line], Montreal, UNESCO Institute for Statistics, 33 p. [retrieved on September 2, 2011]. < http://www.uis.unesco.org/Library/Documents/UIS_strategyreport_2008-2013_en.pdf >.

UNESCO-UIS (2008). *Global Education Digest 2008: Comparing Education Statistics across the World*, Montreal, UNESCO Institute for Statistics, 295 p.

UNESCO-UIS (2011a). *Global Education Digest 2011: Comparing Education Statistics across the World*, Montreal, UNESCO Institute for Statistics, 308 p.

UNESCO-UIS (2011b). *Survey 2011 Data Collection on Education Statistics – Instruction Manual for Completing the Questionnaires on Statistics of Education* [on-line], Montreal, UNESCO Institute for Statistics [retrieved on December 15, 2011], 23 p. < http://www.uis.unesco.org/UISQuestionnaires/Documents/UIS_E_2011M_EN.pdf >.

UNESCO-UIS (2011c). « Country and Regional Profiles », in *Data Centre, Profiles* [on-line], Montreal, UNESCO Institute for Statistics [retrieved on September 5, 2011]. < http://stats.uis.unesco.org/unesco/TableViewer/document.aspx?ReportId=198 >.

UNESCO-UIS Data Center [on-line database], Montreal, UNESCO Institute for Statistics [retrieved on June 4, 2011]. < http://stats.uis.unesco.org >.

WAGNER, Daniel A. (2010). « Quality of Education, comparability, and assessment choice in developing countries », *Compare: A Journal of Comparative and International Education*, Vol. 40, No. 6, p. 741-760.

WALBERG, Herbert and Guoxiong ZHANG (1998). « Analyzing the OECD Indicators Model », *Comparative Education*, Vol. 34, No. 1, p. 55-70.

WALLGREN, Anders, and Britt WALLGREN (2007). *Register-based statistics: administrative data for statistical purposes*, Chichester (England), J. Wiley, 258 p.

WGI data 2011 update [on-line], the World Governance Indicator project, Macroeconomics and Growth Team, Development Research Group, Washington, D.C., World Bank [retrieved on October 14, 2011]. < www.govindicators.org >.

WORLD BANK (2011a). «Country Profiles», in *Education, Data & Statistics, EdStats* [on-line], Washington D.C., World Bank [retrieved on October 10, 2011].
< http://go.worldbank.org/X1LS46NUJ0 >.

WORLD BANK (2011b). « Note on the Statistical Capacity Indicator » [on-line], Washington D.C., Statistical Development and Partnership Team, Development Data Group, World Bank [retrieved on October 21, 2011].
< http://siteresources.worldbank.org/EXTWBDEBTSTA/Resources/3561369-1255619840053/Note_on_Statistical_Capacity_Indicator_2009_BBSC.pdf >.

World Bank Data Catalogue [on-line database], World Development Indicators, Washington, D.C., World Bank [retrieved on November 24, 2011]. < http://data.worldbank.org/ >.