Submission for the 2012 IAOS Prize for Young Statisticians

Title of the paper:

Exploratory analysis of the patterns of missing data in the UIS education database

Name:

Miguel Ibáñez Salinas

Title:

Statistical assistant – Education indicators and data analysis unit

E-mail address:

ibanez_m@ymail.com

m.ibanez-salinas@unesco.org

Employing official statistical agency:

Institute for Statistics – UNESCO

Note: This paper is based on my M.Sc. thesis (HEC Montreal), currently under revision. My thesis supervisor was Dr. Jean-Francois Plante.

Exploratory analysis of the patterns of missing data in the UIS education database

Abstract:

A crucial quality dimension of the UIS education database is its capacity to present complete information on national education systems. This exploratory study examines this longitudinal database from the point of view of the missing values and completeness. This seems to be the first study conducted on this critical matter.

Several analytical tools are used in this research (e.g. binary factors analysis, cluster analysis, multinomial random effects logistic regression), a fact that reflects the complex nature of the patterns of education statistics production. A structure of 5 dimensions or factors (linked to specific parts of the UIS data collection questionnaires) that describes the manner in which countries' education data are produced is discussed. This structure proved valuable in the classification of countries in 5 clusters, where each cluster can be linked to the capacity of countries to produce/report data. The behaviour in time of these factors was also analyzed; the results showed that production of detailed statistics in primary/secondary and tertiary education statistics are decreasing in time. Lastly, evidence about the positive link between improvements in governance indicators and the increase in the production of education statistics is presented.

Exploratory analysis of the patterns of missing data in the UIS education database

Introduction

The UNESCO-UIS education database can be considered as one of the most comprehensive in the world, containing a wide–range of comparable statistics for over 200 countries that governments, international development organizations and researchers use for monitoring education worldwide. For the UNESCO Institute for Statistics (UIS), producing internationally comparable education statistics is a challenging endeavour from both the technical and the political point of view. Nevertheless, the benefits that evidence-based education policies and well-informed citizens have in society are considerable. As a consequence, for the UIS, increasing the quality of its education database is of critical.

Completeness, a critical quality dimension, has been generally defined as the degree or level to which the data are of sufficient breadth, depth and scope for the required functions or tasks (Batini and Scannapieco, 2006). In this regard, the concepts of completeness and "missingness" are intrinsically related. Missing values in the UIS education database can prevent analysts from publishing important indicators, such as the regional averages of out-of-school children or the projections on the number of teachers needed in the world. Moreover, every missing value in the education database increases the number of corrective actions, such as estimations and validation of secondary data, which in turn increases the costs related to data collection.

This paper presents the results of an exploratory study on the different aspects of completeness and missing values across time in the UIS education database. The following discussions are centered on the proposition of a scale for capturing the variability in the production of education statistics (45 items across 5 dimensions), the classification of countries based on their responses as captured by the proposed scale, and the assessment of change in the responses in time and their positive relationship to governance.

Understanding the patterns of completeness/missingness can help to identify countries with reporting problems, and to resolve chronic problems in data collection. At this time, there is no publicly available study on patterns or trends of missing values in the UIS education database.

Data

The present research focuses on the international education dataset (UNESCO-UIS Data Center), which include both "raw data" and indicators.

Raw data are standardized measures of diverse aspects of national education systems, and are the base for calculating indicators. These data are collected annually by the UNESCO education survey, which includes three questionnaires: statistics on education - pre-primary, primary, secondary and post-secondary non-tertiary education (A); statistics on educational finance and expenditure (B); and statistics on tertiary education (C). The UIS Data Center only makes available a partial set of all the collected data.

The education indicators calculated by the UIS are also based on population and economic data produced by the UN Population Division and the World Bank. This study is based on the entire UIS education database, accessed in May 2011, from academic years 1999 to 2008 (2009 and 2010 were still being collected during the writing of this paper).

Each entry (data point) in the database consists of one value symbol or number and, in certain cases, one qualifier (metadata). Apart from the 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: missing value (**...**), magnitude nil or negligible (-) and not applicable (.). Moreover, the value of a data point, if it exists, can be qualified as observed (no symbol added), as a national estimation (*) or as a UIS estimation (**).

The information conveyed in the original series of datasets were transformed into matrices of binary data (response matrices) such that the corresponding binary data point will be 1 when the original element is a numeric value, negligible or not applicable (the metadata information is disregarded) and 0 when the original element 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 any data submitted by countries.

Analysis and results

The present study can be divided in three stages:

1) Factor analysis: the identification of a possible structure of responses in the UIS education database.
2) Cluster analysis: the classification of countries by clusters based on the scale proposed in stage 1 and the interpretation of each cluster.
3) Longitudinal multinomial logistic regression: the characterization of changes in the responses across time on 5 dimensions (as described in the first stage), and the study of the relationship in time of the measures of governance and the changes on responses.

1) Factor analysis

In this section, the major assumption is that an underlying structure exists in the manner in which data points are available (or missing) in the UIS education database. The main output from factor analyses is the proposition of a scale (45 items in 5 dimensions) that represents the underlying structure of responses.

Methods and results

Response matrices are sets of binary data; however, factor analysis is designed for quantitative variables. To circumvent this issue, the corresponding tetrachoric correlations matrices were used for analysis [%POLYCHOR macro (SAS Institute Inc., 2005)]. 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" [Larocque (2006) : 76].

Factor analysis was applied to the tetrachoric correlations from the entire response matrix (522 variables) and from a 45-item sample (the variables most frequently used in international reports) and the results regarding the retained structure, the overall fit and the behaviour of the variables between these datasets were analyzed and compared. The complete dataset would have a subject to item ratio of 1:2.5 (209 countries and 522 variables), well below the recommended 5:1 (Larocque 2006), while the 45-item dataset would represent better subject to item ratio (4.6:1).

The results from the factor analysis applied to the tetrachoric correlations from the complete response matrix for the years 2005-2008 [using both varimax (orthogonal) and oblimin (oblique) rotations] suggests a solution with 5 components. This solution is stable across the years, it has a straightforward interpretation and on average it represents 88% of the total variance each year for data corresponding to academic years 2005-2008. Since the idea that factors related to the

production of educational statistics are correlated seems plausible (the same respondents could generate financial and enrolment data), the oblique rotation will be used in the analysis.

Factor analysis applied to the tetrachoric correlation from the selected 45 variables also points to a solution of 5 dimensions very similar to the solution from the entire response matrices; nevertheless, further improvements regarding item loadings were needed in order to obtain a reliably scale. From this group, nine variables were retired from analysis due to problems with loadings (high in more than one dimension or low loadings in all dimensions). The retired variables were replaced by other variables from the 522-variable pool based on loadings and interpretation.

Taking into account the final set of items, a 5-dimension solution represents, on average, 90% of the variability each year (2005-2008). The communalities are generally high, the residual correlations are low, and most of the time loadings 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, with only a few variables not loading high in the hypothesized dimension. Table 1 presents a brief interpretation of each dimension.

Table 1. Brief description of the 5 dimensions from the 45-item scale

Dimension	# items	Definition of the dimension	UIS education quest.	Questionnaire subject
Factor 1	12	Detailed enrolment/repeaters statistics of primary and secondary E.g.: transition rate, survival rate (grade 5), gross intake ratio (last grade), repeaters, etc.	A	Pre-primary, primary and secondary statistics
Factor 2	8	General raw data enrolment/teaching statistics of primary and secondary E.g.: % trained teachers (total, female), pupil-teacher ratio, etc.	A	Pre-primary, primary and secondary statistics
Factor 3	10	Net enrolment rate/gross enrolment rate/children out-of-school statistics (primary and secondary)	A	Pre-primary, primary and secondary statistics
Factor 4	10	Educational expenditure E.g.: public expenditure on education as % of GDP, distribution of public current expenditure on education by level, etc.	B	Education finance – all education levels
Factor 5	8	Tertiary education statistics E.g.: gross enrolment ratio, inbound mobility ratio, etc.	C	Tertiary education statistics

Nevertheless, the matrix of partial correlations controlling by factors, as in the case of the factor analysis of the complete dataset, has values greater than 1. The overall root mean square off-diagonal partials for this model is 0.6155. For a well-fitted model, this value is expected to be close to 0. 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 re-analyzed. The results, including dimensions and loadings, are very similar to those obtained through factor analysis on the tetrachoric correlation matrix without smoothing. The partial correlations for the model related to the smoothed tetrachoric

correlation matrix are all below 1. The overall root mean square off-diagonal partials for this model is 0.2699, value closer to 0 than in the case of the model with tetrachoric correlation matrix.

2) Cluster analysis

This section presents the results of the classification of countries based on their scores[1] on 5 dimensions related to the response structure for the academic years 1999-2008. The retained solution of 5 clusters is consistent across several years; furthermore, its interpretation is direct and allows a comprehensive look at the need for education statistics capacity building in the world.

Methods and results

First, the response matrices from the dataset of 45 items for each academic year were transformed into matrices of distance (dissimilitude). The Dmatch method from SAS' procedure distance (SAS Institute Inc., 2009) was used to convert the simply matching coefficient – a type of association measure – into a Euclidian distance. Then, the resulting matrices of distance were used as input for the cluster analysis – a hierarchical clustering procedure using Ward's method.

From the analysis of the dendrograms, it was noted that a reasonable choice for the number of clusters falls in the range of 3 to 7. The examination of results, while taking into account the need for a small number of groups that can help with the efficient description of countries, pointed to the choice of 5 clusters each year.

The interpretation of the classification solution is immediate for years 1999, 2002 and 2004-2008. The interpretation of the 5 clusters is as follows:

- Cluster 1: composed of 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 1). On average, this group includes 59 countries (representing 28% of the total of 209 countries).
- Cluster 2: similar to Cluster 1, but with the low response rate for Factor 4 (less than 0.20 across years), that is the dimension related to the availability of education finance's indicators (see Figure 2). On average, 55 countries (26%) are found in this group.
- Cluster 3: composed of 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 – less than 0.15 across years - in Factor 4 (education finances) and 5 (tertiary education statistics) (see Figure 3). On average, 30 countries (30%) are found in this group.
- Cluster 4: 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 4). On 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. In average, 23 countries (11%) are found in this group.

[1] The score on a factor for a given country and year can be defined as the number of items belonging to the factor available in the UIS dataset, divided by the total of items in the respective factor.

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

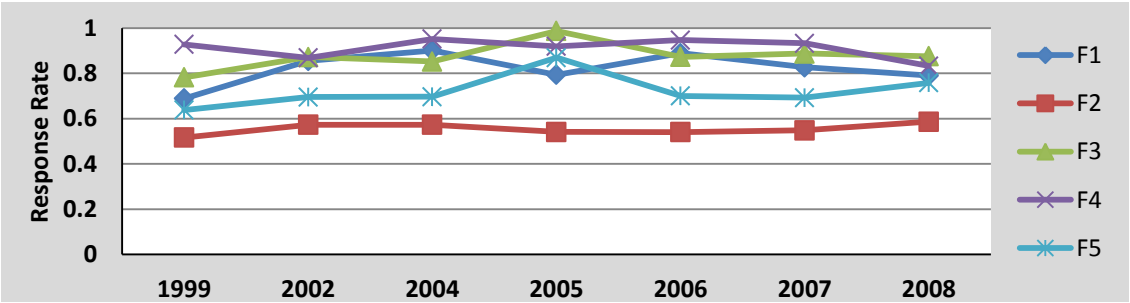


Figure 2. Evolution of the average response rate by factor – CLUSTER 2

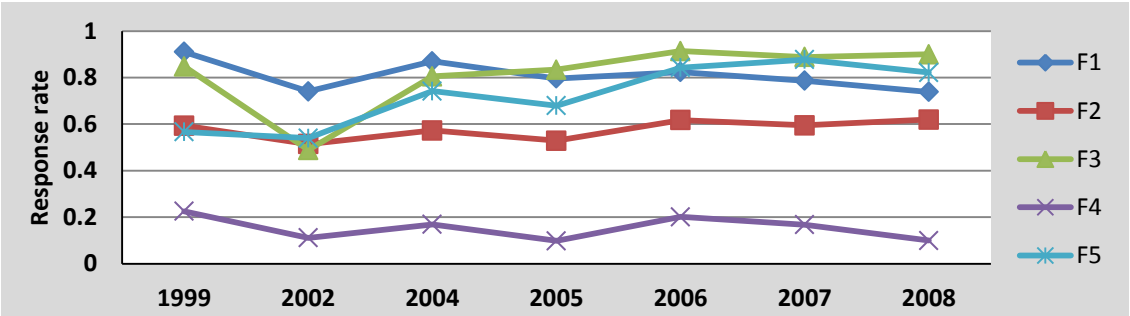


Figure 3. Evolution of the average response rate by factor – CLUSTER 3

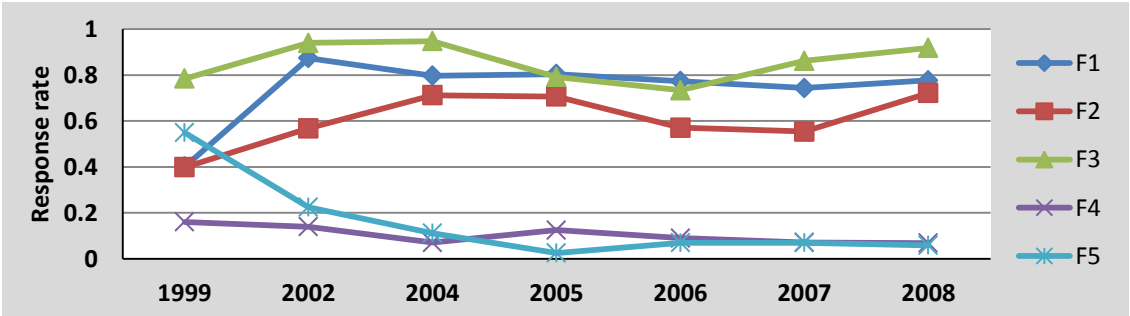
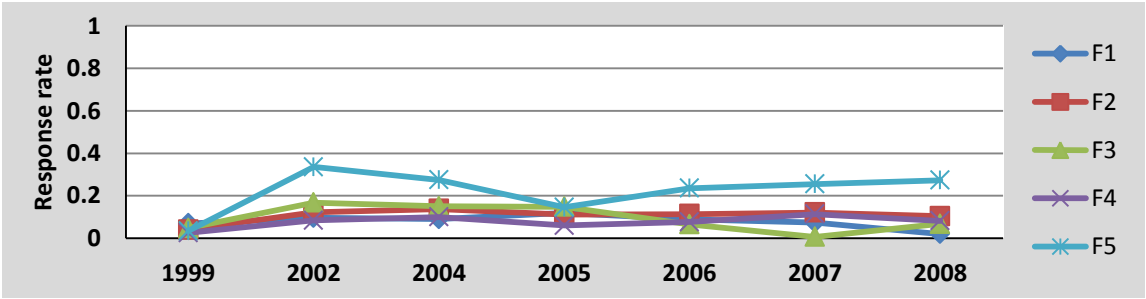


Figure 4. Evolution of the average response rate by factor – CLUSTER 4



Regarding cluster membership, Table 2 presents a list of countries that were classified in a cluster at least 5 times in the years of analysis. In total, we find that 95 countries fulfil this condition. It can be noted that the most populated groups are Cluster 1 (40 constant members) and Cluster 4 (27 constant members).

Table 2. 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**, 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**.

We could also consider clusters as states. In this regard, Table 3 presents a matrix of transition probability for the "average country". It can be noted that countries classified in Clusters 1, 2 and 4 have at least a 50% chance of being classified in the same cluster in a consecutive year. The high probability of remaining in the same cluster for countries in Clusters 1 and 4 (61% and 67.4% respectively) are not surprising: countries in Cluster 1 exhibit consistency in their membership to Cluster 1 due to a robust education statistical capacity, while countries in Cluster 4 may not have the necessary capacity in the short or long term.

Table 3. Matrix of transition probabilities for Cluster 1 to 5 (2004-2008)

		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**

3) Longitudinal multinomial logistic regression

This section presents the results of the study of i) the evolution of the production of statistics around the world through time and ii) the link between governance and statistical capacity related to education statistics. The evidence suggests that tertiary education data and detailed data related to enrolment/repeaters in primary and secondary available in the UIS education database are decreasing in time. The evidence also indicates that there is a significant positive relationship between governance and the availability of education data.

Methods and results

The dependent variables are the countries' score across years on each of the 5 dimensions of the 45-item scale related to the structure of responses in the education database.

The measures chosen to represent governance at national levels come from the Worldwide Governance Indicators (WGI) project, developed by the World Bank (Kaufmann *et al*., 2010; WGI data 2011 update). The WGI project presents measures on three dimensions of governance (roughly described as government monitoring, quality of policy making and respect for institutions) with two measurements for each dimension for over 200 countries for years 1996, 1998, 2000 and 2002-2010. Governance indicators are the result of the standardization and summarization of many perception-based governance data sources across the world. These indicators are reported in the standardized form (mean 0 and standard deviation of 1 across countries per year), and their values range approximately from -2.5 (the worst case) to 2.5 (the best case).

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 [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).

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. The approach of the present study was to see the scores as categories of response, and 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. 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 capita and Total Population were previously transformed to a logarithmic scale.

Regarding the independent variables, four models will be analyzed:

- Full Model: includes all World Government Indicators (WGI) and the control variables.
- Model A: includes governance 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 governance indicators related to government's capacity for efficient policy's implementation (Government effectiveness and Regulatory quality) and the control variables.
- Model C: includes governance indicators related to citizens' respect to economic and social institutions (Rule of Law and Control of corruption) and the control variables.
- Model D: includes only random intercepts and slopes.

SAS' procedure GLIMMIX was used for all models.

In the Full Model, which includes all six governance indicators, the variance inflation factors (VIF) of governance indicators are rather large, ranging from 3 to 20, which denote a possible problem of multi-collinearity. The VIFs 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 as a reference.

The first question, 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 (see Table 4) and Factor 5 (see Table 5), which are related to detailed enrolment statistics for primary/secondary/post-secondary, and to 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 (not shown in this paper). The estimates of slopes are identical in signs and similar in values when comparing the Full Model to Models A, B, C and D.

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.).

The second question, to determine if changes in governance are linked to changes in 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 have a 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.

The positive effect of good governance on the production of education data 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.

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 tests of the random intercept and slopes reject the null hypothesis that they are zero, indicating that a multilevel model is appropriate to the characteristics of the dataset.

Table 4. 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

Conclusion

A 5-dimension scale of 45 items was proposed in order to capture and efficiently manage the variability of responses in the education database. This proposed structure allowed a 5-cluster classification of countries. These structures are robust in time and could be the basis for future diagnostic analyses.

A relevant matter for education analysts, donors, data users and governments is the assessment of the evolution in time of the production of international education statistics. It was concluded that the production of detailed statistics on enrolment/teaching staff related to primary and secondary education (Factor 1) as well as the production of tertiary education statistics (Factor 5) are decreasing in time. 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 rate – are necessary. Further studies at the field level are required in order to understand the problems that countries face when reporting these data. Data validation, estimation and field work in general may become a priority when dealing with decreasing data. It may be necessary to find cost-efficient procedures for data processing as well as effective strategies to encourage countries to report complete data to the UIS.

The longitudinal study of factor scores also points out that increases in governance, as measured by the World Governance Indicators, are positively linked to increases in production of education statistics. This result illustrates the positive relationship between statistics and governance, which is one of the ultimate goals of statistical production (encouraging evidence-based policies).

Table 5. 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

References

- BATINI, Carlo and Monica SCANNAPIECO (2006). *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*, New Jersey, Springer-Verlag, 262 p.
- 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.
- LAROCQUE, Denis (2007). *Analyse multidimensionnelle - Recueil 6602F*, Montréal, HEC Montréal, 409 p.
- 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 >.
- 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.
- 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 >.
- *UNESCO-UIS Data Center* [on-line database], Montreal, UNESCO Institute for Statistics [retrieved on June 4, 2011]. < http://stats.uis.unesco.org >.
- *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 Data Catalogue* [on-line database], World Development Indicators, Washington, D.C., World Bank [retrieved on November 24, 2011]. < http://data.worldbank.org/ >.