# Enrolment ratios and related puzzles in developing countries: Approaches for interrogating the data drawing from the case of South Africa

Martin Gustafsson [*]

Department of Economics, University of Stellenbosch, Private Bag X1, 7602 Matieland, South Africa

## ARTICLE INFO

## ABSTRACT

Available developing country enrolment ratios are frequently inaccurate. This results in poorly informed policy discourses. Underlying data problems can undermine funding systems. The magnitude of the problem is described, partly through reference to UNESCO and UNICEF ratios. Steps for remedying the situation within one country are proposed, on the basis of an existing South African study, and some new data analysis of South African, Brazilian, and international data. Remedies must be sensitive to local contexts, but proposed techniques for detecting fraud in the enrolment microdata, and comparing age structures across different datasets, seem relevant for different contexts.

© 2015 Elsevier Ltd. All rights reserved.

## 1. Introduction

The discourse on progress in education systems is extremely sensitive to the quality of data on matters such as participation and learning achievement. This article, which focuses on participation, seeks to highlight that the available statistics are too often misleading, to the detriment of the policy discourse and operational processes such as school funding. It draws from a specific South African study where enrolment ratios were interrogated (Gustafsson, 2012), as well as further analysis of data from South Africa and beyond, in proposing a series of steps which researchers can take to fix poor enrolment ratio statistics.

Section 2 draws from the limited literature on the problem of poor data quality in the social sciences in arguing that there is too little healthy scepticism about official data amongst researchers. Suggestions made in the literature to improve the situation are discussed. Previous work by Stukel and Feroz-Zada (2010) on the reliability of existing enrolment ratios, work on which this article builds, is summarised. Section 3 presents an analysis of large discrepancies between UNESCO's country-level enrolment ratios, which are based on official enrolment and population estimates, and UNICEF's ratios, which use sample-based household data.

Section 4 explains seven steps for problem detection and remediation. It includes proposals for bringing together attendance

and attainment values from the household data in order to test the validity of the former. It also uncovers unexpected rounding patterns in both South Africa's and Brazil's school census data. Overall, the finding is that the quality of school census data is improving, at least in South Africa and Brazil, that there appears to be less cause for optimism when it comes to official population figures, and that sample-based household data remain an important and relatively reliable source which can be used to verify and even adjust the education and population census totals. Section 5 concludes the article.

## 2. Attention devoted to data problems in the literature

Governments are major producers of both census and sample-based data. Moreover, governments often work closely with institutions collecting international data, such the sample-based testing data of programmes such as PISA[1] and SACMEQ.[2] Education researchers make extensive use of these data. To illustrate, of the 29 articles published in the first half of 2013 in the International Journal of Educational Development, eight included statistics derived from data produced in some way by governments, and in the case of six articles new analysis of the microdata was conducted. A further eight articles made use of non-government data, largely smaller collections produced by research organisations for specific research

---

* Correspondence to: PO Box 8918, Pretoria 0001, South Africa. Tel.: +27 82 853 5951.
E-mail address: mgustafsson@sun.ac.za

[1] Programme for International Student Assessment.
[2] Southern and Eastern Africa Consortium for Monitoring Educational Quality.

purposes. The current article deals with problems and solutions relating largely to government data.

Problems in the education data do not only contaminate the policy discourse in education. Many disagreements amongst economists about the precise role of educational progress in economic development have their roots in differences of opinion over how to use household and school census data to construct measures of educational attainment (Gustafsson, 2014: 50). Krueger and Lindahl (2001) have argued that problems with country-level enrolment and attainment data are part of the reason why the role of education in economic development has been insufficiently appreciated.

Jerven (2013) offers a rare in-depth assessment of the problem of poor quality official data and suggestions on how the data could be improved. Whilst his book's main focus is on data needed to measure the economy, and his scope is limited to Africa, the points he makes are pertinent to the arena of education research, in particular in developing country contexts.

Four suggestions made by Jerven (2013) stand out as relevant for the present discussion. One is that human and institutional capacity to collect and analyse data needs strengthening. The challenge is not just to build technical capacity, but also to combat what Inbar (1996: 92) would refer to as ritualism in governments (or international agencies), whereby the act of merely collecting data and publishing statistics becomes a virtuous ritual in itself, regardless of what the data actually tell or their reliability. Data, especially if not interrogated, can easily become a cornerstone of what Pritchett et al. (2012) would call the problem of mimicry in government, or the pretence that the state is working as it should, when in fact it is not. The risk that abuses will arise, almost deliberately, is wryly acknowledged in the following extract from a 1990 manual of the International Institute for Educational Planning (Ross and Mälck, 1990: 100):

In essence, improving the collection, preparation, and analysis of data requires attention to detail. That is, there are no "shortcuts" to achieving high standards in these areas. However, there are many easy ways to ensure that standards are low and that the results of the data analyses are meaningless.

The implied warning seems as relevant a quarter of a century later, despite major advances in the technology surrounding data collection and analysis.

Jerven's (2013) second suggestion is that accountability bodies and data users should insist more strongly on truly useful technical documentation accompanying datasets, documentation which goes beyond the usual limited focus on matters such as the sampling methodology and a codebook. Porta and Arcia (2011: 16), authors of a World Bank guide to improving education data, are right to insist on methodological papers and adherence to international standards. However, it should perhaps be made clearer what good methodological papers should look like. For one, they need to bring to a wider audience knowledge about the limitations of available data which more experienced users around the world have been aware of for a long time. Key limitations that stand out are inaccuracies in enrolment breakdowns by age, numbers of grade repeaters and (a key concern in this paper) inconsistencies across different datasets, which, according to the manuals, ought to be combined to calculate ratios.

The notion that there are highly developed 'international standards' which analysts can rely on is a questionable one. Jerven (2013: xvii) is particularly critical of the custodians of international data standards, specifically the IMF and the World Bank, which are seen to be more interested in 'maintaining the official validity of the numbers they use' than worrying about data quality or being transparent about their methods. Whilst the UNESCO Institute for Statistics (UIS) has done important work in defining, for instance,

education levels, there is still much work to do. UIS manuals on how to calculate education indicators tend to be theoretical, without assessments of what occurs when one actually implements these methods using available data. An important exception here is the report by Stukel and Feroz-Zada (2010), which was commissioned by the UIS. Of course it is impossible for the UIS to take into account the data idiosyncrasies of every country from which it gathers data. However, the detection of patterns within the country-level UIS data done for the current article, as well as the South Africa case study discussed below, are partly intended as indications of work that could be done more frequently and systematically. Many developing countries remain highly reliant on bodies such as the UIS in deciding on how to collect and process education data, implying that a more rigorous conversation between the global and national levels on data quality, in particular where issues are common across many countries, could lead to better quality data over time.

How the matter of school dropouts was dealt with in the South African policy discourse in recent years provides an indication of the cost of not paying attention to having quality data, or at least documentation on the limitations of the available data. The South African Grade 1 school dropout rate was reported by UNESCO (2007: 306) to be 10% in 2004. Dropout rates for South Africa, as for other countries, were calculated by the UIS using a standard formula. To those knowledgeable about, firstly, the reliance of this indicator on accurate grade repetition data, when using the UNESCO formula, secondly, incentives existing for schools to under-report grade repetition and, thirdly, what the household data were saying about school participation, the 10% value was clearly the result of an error. However, as there was no published critique of the 10% statistic in question, and as one was dealing with a UNESCO publication, it seemed reasonable for the South African Human Rights Commission to investigate the matter. In fact, the Commission, on the basis of another poor analysis, arrived at the conclusion that as many as 26% of Grades 1–3 children were dropping out of school (South Africa Department of Education, 2008: xiii, 10). This led the minister of education to launch an official investigation and initiate more in-depth analysis, which confirmed that the dropout rate for Grade 1 was around 0.2%, certainly not a national crisis. The cost of allowing a relatively basic data error to create an imaginary crisis, which then takes up a substantial share of the policy discourse, is large. As Crouch (2005: 3) suggests, 'distracting debates' are one factor which has prevented more focussed attention on truly serious problems in South Africa, such as inequalities in learning outcomes.

A third suggestion put forward by Jerven (2013) is that researchers need to strengthen their ability to interrogate numbers. 'Scholars need to ask themselves the same questions when confronted with numbers as they would when they are confronted with research findings: How did you arrive at this result?' (Jerven, 2013: 120). This is not the same as calling for researchers to dust off their statistics textbooks. Interrogating numbers involves understanding the history of data collections better, examining original questionnaires, thinking through the incentives affecting everyone from respondents to the compilers of final metadata reports, searching for suspicious patterns in single datasets, and comparing patterns seen in different datasets.

Fourthly, Jerven (2013: xii) argues that scholars need to produce more 'ethnographies' of data, or formal and published inquiries into the data that underpin much of our knowledge and research. Whilst sweeping statements about data being bad are common, what is missing are more inquiries into *how bad* the data are. In education, there are encouraging signs that at least some probing below the surface of widely quoted country-level test score averages is occurring (see for instance the work of Jerrim (2013a,b) on the PISA data). However, there remain important