The Numbers Game: Suggestions for Improving School Education Data

KIRAN BHATTY

In the context of the declining quality of public education, governance has emerged as an important explanatory variable, quite distinct from the education variables more commonly cited, such as teaching and learning practices or curriculum and textbook quality. An important component of the governance architecture in any sector is its information and data regime, as all aspects of monitoring, planning and policymaking are dependent on it. A look at the data system in the education sector in India reveals that there is much amiss at all levels of data collection and use.

This is not to deny that compared to a couple of decades ago, considerable energy and investment have gone into building a regular school-based decentralized data collection system in India. This District Information System for Education (DISE), set up after Sarva Shiksha Abhiyan (SSA) was launched in 2001, and now called Unified-DISE (U-DISE),[1] collects data from 1.5 million schools (government and private) and provides report cards up to the secondary stage for every state, district and school. It is remarkable that this data is compiled and School Report Cards prepared and uploaded on the website on an annual basis. Education data from households is also being collected by Panchayats and compiled annually in Village Education Registers. A few states have supplemented this with data from Child Tracking Surveys, which enumerate the population of school-going children. In addition, the Ministry of Human Resource Development (MoHRD) commissioned three rounds of household surveys in 2006, 2009 and 2014. The SRI-IMRB surveys, as they are called, collect information on children in the age group 6-13 years who are out of school. Other large household data sets have emerged too, in addition to the National Sample Survey (NSS) and Census, such as the National Council of Applied Economic Research's (NCAER) Indian Human Development Survey (IHDS-I, 2004-5 and IHDS-II,

2010-11), the Annual Status of Education Reports (ASER) since 2005, and now the Socio-Economic Caste Census (SECC). All of them provide data on education indicators and school participation in some form.

However, in the midst of this 'feast' of data sources, we get varied, often contradictory evidence on basic indicators such as the proportion of children out of school, the extent of improvement in retention levels, the learning outcomes and the quality of education. Even in areas of education finance, such as teacher appointments and salaries, we do not have an authentic database. Hence, despite the fact that the coverage and scope of data collection by the government has increased enormously with many more indicators added, nagging questions remain about the quality, utility and purpose of the data, with obvious implications for planning and policymaking. Further, with multiple sources of data – both governmental and non-government – in operation, data neutrality also cannot be assumed.

This paper highlights the methodological as well as administrative anomalies in the system, and points to the need for greater decentralized management of data as well as collaboration across agencies for purposes of standardizing definitions and methods of estimation. It further emphasizes the need for public verification of data to ensure authenticity as well as validation across sources to reduce bias.

Methodological Discrepancies

Definitions and Methods of Estimation

The methodological difficulties begin with the range of definitions and methods of estimation used for important indicators by different government and non-government agencies collecting data. For instance, estimates for out-of-school-children (OOSC),[2] all collected through household surveys, are based on different 'questions' asked by investigators employed by each source. The NSS, for example, asks, 'How many children are currently attending school?', while the Census enumerators ask questions related to 'status of attendance in an educational institution'. The MoHRD survey, on the other hand, claims to follow both the sampling and methodology used by the NSS, and yet arrives at vastly different results. The NSS and MoHRD surveys, which are based on a sample, then extrapolate from their figures the proportion of children that are out of school as a percentage of the population of children in that age group. Using this method, the NSS 71st round (2014) has pegged the figure at a little less than 10% of the child population, amounting to nearly 20 million children, while the MoHRD (SRI-IMRB, 2014) estimates put it at 3% and thus approximately 3 million! The 2011 Census, on the other hand, suggests that more than 15% children in the same age group do not go to school, thus giving us a wildly differing figure of almost 40 million.

Similarly, the figure for the total number of teachers in a school turns out to be not as simple a statistic as it sounds, with teachers being routinely sent on deputation to other schools.[3] Thus, it is unclear whether a teacher who is on deputation from another school is to be counted in her current position or in her original school; or does she end up being counted in both? Similarly, information on the employment status of teachers has only two categories in the DISE format – regular and contract – whereas multiple categories that do not fit precisely into these categories also exist (voluntary, assistant, etc.), resulting in highly inaccurate data being collected on such an important indicator. Other gaps in the data collected include: information on salaries paid out by each state to the different categories of teachers and measures of learning outcomes on a regular basis. The problems are compounded by the fact that formats for collecting data are designed centrally and do not take into account local specificities; nor are teachers – often the primary data enumerators – adequately trained to fill the formats.

Validation and Verification of Data

Another aspect of data credibility that has proved to be a weak link in the data collection process is verification and validation of data. While the rules for DISE dictate that 10% of the sample be randomly cross-checked, DISE itself is unable to verify that this process is either regularly or adequately carried out, due to lack of capacities available at the frontline for the process. In addition, the education departments

ignore the evidence presented by other government or non-government sources to validate and thus improve the credibility of their data. Data validation faces some mundane difficulties as well, related to different methods and time periods used for estimating different indicators by the agencies that collect data. For instance, the Right to Education (RTE) Act talks about children between 6 to 14 years age, but practically all data agencies (except those under MoHRD) use different age groups when compiling education data, making comparison quite difficult. Similarly, the dates and periodicity of data collection vary across sources. ASER is an annual survey; NFHS followed a six-yearly pattern initially but has now slipped to 10 years since the last survey. IHDS thus far has maintained a gap of six years between its two successive surveys. While NFHS-3 and IHDS-1 roughly cover the same period (2004-5 and 2005-6), neither corresponds to the Census dates, but IHDS-2 (2011-12) does. NSS also follows a different time period for its education surveys.

Administrative Anomalies

The Purpose of Generating Data

Different agencies plan their data collection for different (and specific) purposes, and not necessarily for planning or monitoring education and hence for education policy. For example, the education rounds of NSS are part of the survey on social consumption, which in turn seeks to assess the benefits derived by various sections of society from public expenditure incurred by the government.[4] The population census, on the other hand, is the primary source of basic national population data required for administrative purposes and for different aspects of economic and social research and planning.[5] The non-government sources also have unique purposes in mind, again not necessarily with education as the primary objective. Thus, NFHS is essentially a health and nutrition survey that also collects data on select education parameters. Similarly, IHDS is geared towards the larger goals of human development and poverty, especially the links between education, skills and livelihood. Only ASER is solely dedicated to education, specifically learning levels. However, it does not tell us how the levels of learning vary with student enrolment or attendance, or any household factor.

What is more surprising is that even the data collected by MoHRD and state education departments, though admittedly for the purpose of monitoring and planning education, is not geared towards policy goals. Instead, data collection and analysis are guided by their use in taking stock of the provisioning of schools, rather than as a mirror of their functioning. Unsurprisingly, therefore, school surveys focus on collecting information related to (i) broad indicators of infrastructure and teacher availability; and (ii) student enrolment and distribution of incentives. Both these sets of data showcase administrative efforts rather than education progress. Even the household survey (MoHRD's SRI-IMRB) is used only for estimating OOSC. No effort is made to use disaggregated data to understand the problems of specific groups of children or schools.

A second conundrum associated with the purpose and use of education data relates to the fact that planning and policymaking are extremely centralized processes. Thus, data – however collected – plays a limited role in the planning and policy processes. For instance, the Project Approval Board at the MoHRD that approves

annual plans and budgets (AWP&Bs) for the states does so on the basis of the finances allocated to it by the Ministry of Finance and the norms of expenditure specified by the central ministry (MoHRD). While the AWP&B for a state reflects the needs of the state, eventual allocations differ widely from it, as they are based on what is made available by the Ministry of Finance through processes that do not involve the education sector. Of course, state plans are themselves based on a process of aggregation that does not involve a genuine decentralized planning process. This is evident from the fact that dissemination strategies are not aligned with the goals of decentralized planning, as data is largely unavailable in usable form at the local or school level. In fact, local data management systems are virtually non-existent, putting paid to the idea of decentralized planning. Thus, while it is true that schools are now asked to prepare their plans through the School Management Committees, in fact what is submitted by them are copies of the DISE format – presumably as indicative of the status of schools and thus reflective of their needs! Eventually, therefore, at the district level – and probably also at the state level – DISE data is referred to for determining the state AWB&P.

Limited State Capacity

A second and perhaps overarching problem confronting the data regime in education is that of limited capacities to design, collect, analyse and use data throughout the government structures, from the central to the local. DISE is run almost entirely on the shoulders of data entry operators of the education departments at the district and block levels. Data that is collected from the ground up amounts to a process of simple aggregation resulting in the loss of specifics, such that by the time it reaches the central level, it barely reflects the ground realities and can hardly serve the needs of the people. The aggregation itself is still done manually at the block level in many states with digitalization appearing only at the district level. Further, implicit in the collection process is a conflict of interest, especially with DISE data as it is entirely dependent on formats filled by teachers. It is well established that teachers might be incentivized to represent information in ways that inflate facts, such as student enrolment.[6]

In addition, the departmental staff at the state level have not acquired the understanding, through their own qualifications or through training provided by the government, of the relevance and importance of quality data or its use in the planning or policy process. For instance, innumerable formats are designed for monitoring schools, but none of that data is put to any use.[7] In fact, it is not even referred to in the monitoring or review meetings held at the block and the district. Unfortunately, the personnel involved in collecting and collating that information are themselves unable to gauge its importance as they see it as simply a chore – of 'filling formats'. With the import of the data completely lost on them, they are unable to use it in a constructive fashion, making the entire exercise redundant.

The Way Forward

The new draft National Education Policy, 2019, in recognizing the paucity and limitations of the education data regime, has called for "a major effort" in data collection, analysis and organization. In particular, it proposes the establishment of a new Central Educational Statistics Division (CESD) as an independent and autonomous entity at the NIEPA. It has also suggested the maintenance of a National Repository of Educational Data (NRED) within NIEPA, which will include specific indicators common to under-represented groups (URGs), in an attempt to track their participation and performance. Building a local data base for drop out and out of school children, using social workers to collect information, as opposed to teachers, is another welcome suggestion. Making State Assessment Survey (SAS) results available transparently to parents, teachers, SMCs members and the community could also add to community participation in the learning outcomes of students, as well as validation of the data and accountability to the people.

Key issues that the CESD and related authorities will nevertheless need to deal with are mentioned below:

(i) Improving definitions, standardizing them across sources, and using improved methods of collection and estimation of basic indicators.

(ii) Developing capacities of the data regime and giving a greater role to data users, especially the education officials at different levels of government ranging from the national to the local. Necessary technical skills, if provided, will enable them to be cautious when collecting data, as also to interpret and use it appropriately, such as when making plans.

(iii) Providing support to monitoring agencies, such as the school management committees, school complex management committees, social audit groups, and education researchers to allow them to publically verify data that is officially collected. This requires building a local data management system – at the level of the school or Panchayat or School Complex – that has more than out of school data, as proposed by the NEP, and is publicly available. It would go a long way in facilitating not just local monitoring but also the development of school development plans. In the current situation, the lack of computing facilities at the local level inhibits the maintenance of data, as paper records tend to be poorly maintained and not updated. As a result, even the information that is generated in the school is sent up to the next level for digitization at a higher level where computing facilities are available, at the district or block level, as the case maybe. The digitized information however, does not flow back to the school, for the same reason. As a result, no institutional memory is built up for purposes of tracking change or progress in a school. Ideally the format should be verified, by the parents and larger community, before being sent up to ensure accuracy.

(iv) Reducing bias by validation through the use of multiple data sets. Validation of data against different sources, especially in the case of data used for policy, can ensure that bias is factored in and therefore a more judicious use of data is effected. Multiple data sets have other uses as well. For instance, while any single data set cannot collect information on all relevant issues, data collection is known to be a very expensive and time-consuming process. Thus, information collected by NSS on household expenditures – which demonstrates that 70% of all OOSC in urban areas are concentrated in the lowest quintile, while in rural areas they are in the lowest two quintiles – is relevant information that can and should be used by the education department without having to repeat the exercise. Similarly, NFHS data provides linkages between education participation and family health, also of importance to the education department.

(v) Making better use of data through proactive collaboration of different government and non-government agencies. For instance, if household and school data were available in the same portal, it would maximize their use. Similarly, if the NSS education rounds were better coordinated, along with standardization of definitions of important indicators, it would greatly help in serving the cause of education goals. Streamlining the planning process to enable planning based on decentralized data will go a long way towards improving the use of data at the local level as well as ensuring a more genuine decentralized planning process.

END NOTES

1. U-Dise or Unified-DISE is a database of all students from grades 1 to 12.
2. Non-government sources do not collect information on this variable at the national level.
3. It is common to send a teacher appointed to a particular school to another, if there is a shortage in the other school. While shortages exist in a very large number of schools, such deputation typically takes place if the demand for more teachers is raised loudly enough or the political configuration is such that the school is able to draw a teacher towards their school, typically creating a shortage in the school from which the teacher is deputed!
4. http://mail.mospi.gov.in/.
5. http://censusindia.gov.in/Data_Products/Library/Indian_perceptive_link/Census_Objectives_link/censusobjectives.ht
6. See Bhatty, Saraf and Gupta, 'Out-of-school Children in India: Some Insights into What We Know and What We Don't', *Economic and Political Weekly* 52(49) (2017)
7. See Bhatty and Saraf, 'Does Government's Monitoring of Schools Work?', CPR Working Paper (New Delhi: Centre for Policy Research, 2016).