LETTER

The data quality debate on Indian surveys should be more responsible

Published online first on September 8, 2023. DOI: 10.20529/IJME.2023.052

Recently, the data quality of the National Sample Surveys (NSS) and the National Family Health Surveys (NFHS) has become the centre of discussion [1,2]. Two issues that have been raised include the overestimation of the rural population in these surveys and greater response rates in poorer wealth groups compared to the richer groups. Technically, there are concerns about the generalisability of these surveys. Politically, the argument is that together these issues bias the surveys toward depicting the country as worse off. In other words, the surveys do not capture the growth in urbanisation and accompanying wealth generation that has happened over the recent past.

Debates around data quality are important as these data form the basis for the country's policymaking. However, the ethics of such debates needs to be grounded in evidence and should utilise the appropriate mechanisms of scrutiny for ascertaining the validity of arguments. The current debate is neither evidence-based nor is it framed as rigorous academic debate should be — with far-reaching consequences for public policy and public perceptions. The descent of the discourse into newspaper articles making opposing and somewhat unsubstantiated claims as well as social media spats, is arguably not in good faith.

On the technical front, the issue of rural population overestimation might be more challenging than its current portrayal. First, commenting on overestimation needs a sound standard of comparison, ie, ground truth. Typically, the census would act as the ground truth. However, the last census for India was completed 12 years ago, which makes any comparison difficult. In the absence of data, one can rely on projections. However, the reliability of the projections can be questioned in the same fashion as the survey-based population estimates are being questioned. Hence, anyone concerned with data quality and interested in assessing it should be concerned about the absence of ground truth data, in this case, the census. Second, there is no consensus or method of measurement to determine how much overestimation is tolerable. For instance, in our article, we noted that the difference between the census-based population projections and NSS-based estimates for rural population proportion ranges from 2.57% points to 4.40% points across years. [3]. It is difficult to say whether this difference is alarming or not. There is no threshold as such. Third, the source of the overestimation remains unknown. The debate until now has generated speculation and promoted the opinion divide. However, we are still far from asking and answering what design elements of the surveys are creating this bias.

Another important issue is the difference in response rates across wealth groups. Again, the debate has focused emphatically on the implications of such differential response rates without establishing their existence. We analysed multiple response categories (cooperative and capable to respond, cooperative and not capable, busy, reluctant, and other) across wealth quartiles for eight NSS surveys covering a period from 2011 to 2019 [3]. Specifically, we looked at the differences in the response rate estimates between the richest and poorest quartiles for each response category. We found a positive difference in the proportion of respondents who were cooperative and capable between the richest and poorest quartiles highlighting that the cooperative and capable respondents were more likely to belong to the richest quartile. Our analysis also showed that the percentage point difference in reluctant and busy respondents between the richest and poorest quartiles varied only marginally. Hence, the bias, if any, was negligible and could not skew the findings decisively in any way. It is also important to note that a greater proportion of respondents who were cooperative but not capable of responding belonged to the poorest rather than the richest quartile. This further diminishes the threat of existing surveys being biased towards overestimating the percentage of poor residents.

On the political front, this debate has opened the door to multiple problems with ethical implications. First, it has transferred a supposedly academic discussion to a public platform where the arguments with more popular support are being valued over those with valid content. Second, the debate has actively contributed to diminishing public faith in institutions that are responsible for data generation at a time when denial of data, misinformation, disinformation, and politically charged narratives run high. Finally, a precise diagnosis of the problem and any directions for managing them remain elusive. This makes the debate on data quality issues unproductive and raises concerns about its purpose, to begin with. Bringing a technical debate to a public platform has only resulted in confusing people without providing answers.

Academics and policymakers are specialists whom society trusts. Debates are highly valued in a free and healthy

intellectual culture. Hence, the onus of how to responsibly initiate and conduct such debate lies on the specialists. Acting irresponsibly is a violation of ethics.

Conflicts of interest and funding: None

Siddhesh Zadey *(corresponding author - sidzadey@asarforindia.org), Association for Socially Applicable Research, Pune; Dr. D.Y. Patil Medical College, Hospital, and Research Centre, Pune, Maharashtra, INDIA;* ***Parth Sharma*** *(parth.sharma25@gmail.com), Association for Socially Applicable Research, Pune, Maharashtra; Department of Community Medicine, Maulana Azad Medical College, New Delhi, INDIA;* ***Pushkar Nimkar*** *(pushkarnim@gmail.com), Association for Socially Applicable Research, Pune, Maharashtra, INDIA*

References

1. Ravi S. Shamika Ravi writes: Our national surveys are based on faulty sampling. *The Indian Express*. 2023 Jul 7 [Cited 2023 Aug 27]. Available from: https://indianexpress.com/article/opinion/columns/shamika-ravi-writes-our-national-surveys-are-based-on-faulty-sampling-8799300/
2. Sen P. Pronab Sen responds to Shamika Ravi: No, India's statisticians aren't stupid. *The Indian Express*. 2023 July 11 [Cited 2023 Aug 27]. Available from: https://indianexpress.com/article/opinion/columns/criticism-of-sample-surveys-is-misplaced-their-data-differ-from-census-count-because-definitions-are-different-8822347/
3. Zadey S, Nimkar P, Sharma P. Evidence, not narratives, should guide discussions about statistics. *The Hindu.* 2023 Aug 3 [Cited 2023 Aug 27]. Available from: https://www.thehindu.com/opinion/evidence-not-narratives-should-guide-discussions-about-statistics/article67153810.ece