EAC-PM Working Paper Series
EAC-PM/WP/21/2023

Assessing the National Surveys for its Representativeness

An Analysis of the Data Quality of the National Sample Survey (NSS)

Economic Advisory Council to the PM

July, 2023

Shamika Ravi
Mudit Kapoor
S V Subramanian

Abstract

This paper is a quantitative analysis of the data quality of the National Sample Survey (NSS) in terms of three estimates, (i) the proportion of the rural population, (ii) the proportion of the Scheduled Caste (SC) population, and (iii) the proportion of the working-age (age between 15 and 59 years) population. We follow Meng [2018] to demonstrate that the data defect correlation, a measure of the correlation between the indicator variable, which takes a value of 1 if the population unit is selected and 0 otherwise, and the variable of interest, is significantly high, which warrants a reduction in the bias-adjusted effective sample size from more than 4.5 lakh observations to less than 500 to 5000, a reduction in statistical efficiency ranging from 97% to 99.8%. The paper has implications for surveys that use the same sampling strategy, such as the National Family Health Survey (NFHS) and the Periodic Labour Force Survey (PLFS). We emphasise that increasing the data quantity cannot address data quality issues. On the contrary, it leads to *Big Data Paradox* (Meng, 2018) "The more the data, the surer we fool ourselves."

Assessing the National Surveys for its Representativeness

An Analysis of the Data Quality of the National Sample Survey (NSS)

*****

Shamika Ravi [*]
Mudit Kapoor [†]
S V Subramanian [‡ §]

1. Introduction

Policymakers in India have relied on the National Sample Survey (NSS) of the household to assess the poverty and employment situation at the national and state levels. The NSS is conducted by the Ministry of Statistics and Programme Implementation (MOSPI), Government of India.[1] Estimates from the NSS play an instrumental role in studying the impact of various government policies on poverty and employment and, more generally, are an indicator of the government's overall performance towards the welfare of people, in particular, the socially disadvantaged and the vulnerable, such as the Scheduled Caste (SC). The NSS has also been used to provide estimates of persons with disabilities or assess the provision of drinking water, hygiene, and conditions of the house, or provide estimates of expenditures related to domestic tourism, etc. Given the significance of the NSS in policymaking, it is important to examine its data quality quantitatively, mainly whether it is nationally representative.

To examine the data quality of the NSS, we follow Meng [2018][2], who decomposes the total survey error (TSE), which is the difference between the estimate from the Survey and actual population values, into three distinct components. These are (a) data quantity, which is related to the fraction of the population that is sampled, (b) measure of problem difficulty, which is the variability of the outcome of interest in the population; and (c) data quality, which intuitively is the correlation between the indicator variable, which takes value 1 if the population unit is sampled and 0 otherwise, and the variable of interest.[3–6]

The real challenge with quantifying data quality is that actual population values are not easily and independently available for the same period as the Survey. However, in this note, we overcome this challenge. In particular, we

[*] Member, Economics Advisory Council to the Prime Minister, NITI Aayog, Sansad Marg, New Delhi, 110001, India

[†] Economics and Planning Unit, Indian Statistical Institute, Qutub Institutional Area, New Delhi, 110016, India. Email: mkapoor@isid.ac.in

[‡] Department of Social and Behavioral Sciences, Harvard T. H. Chan School of Public Health, Harvard University, 677 Huntington Avenue, Cambridge, MA, 02115, USA

[§] Harvard Center for Population and Development Studies, Harvard T. H. Chan School of Public Health, Harvard University, 677 Huntington Avenue, Cambridge, MA, 02115, USA

compare two NSS, one for household consumption expenditure[7], which is used for poverty estimates, and the second for employment and unemployment[8]. These surveys were the 68th round of the NSS, conducted between July 2011 and June 2012. For population values, we use the values from the Census conducted in 2011[9].

It is commonly perceived that data quality issues can be addressed by increasing the sample size. However, on the contrary, it has been convincingly shown[3,4] that the bigness of the data in the presence of data defects is subjected to the "*Big data paradox*: the more the data, the surer we fool ourselves."[3] Therefore, it becomes essential to exclusively study the data quality of the NSS, given that its sample size is over 450,000 observations[7,8].

In this note, we study three estimates from the Survey and compare them to the values from the Census. These are the proportion of the rural population, the proportion of the Scheduled Caste (SC) population, and the proportion of the working-age (age between 15 and 59 years) population. In addition, we also study the data quality of the NSS for rural and urban areas separately. For rural and urban areas, we compare the estimates from the Survey with values from the Census for the proportion of the SC population and the proportion of the working-age population. Our choice for these variables is driven by the fact that these variable correlates with measures of poverty, employment, and unemployment. Also, these variables are consistently measured across the Survey and the Census.

To our knowledge, this is the first note that quantifies the data quality of the NSS to assess its national representatives.

2. Methodology for Quantifying the Data Quality

To assess the data quality of surveys, we follow Meng [2018][3]. Let us suppose that the variable of interest that is estimated from the sample survey of n observation is $\overline{G_n}$, while the population value is $\overline{G_N}$, then the total survey error, $\overline{G_n} - \overline{G_N}$, which can be decomposed into three factors: data quality, data quantity, and the problem difficulty.

$$\underbrace{\overline{G_n} - \overline{G_N}}_{Total\,survey\,error} = \underbrace{\rho_{R,G}}_{Data\,Quality} \times \underbrace{\sqrt{\frac{1-f+CV_W^2}{f}}}_{Data\,Quantity} \times \underbrace{\sigma_G}_{Problem\,Difficulty}.$$

Where f is the fraction of the population included in the sample, CV_W^2 is the square of the coefficient of variation (standard deviation/mean) of the sampling weights, and σ_G measures the variability among the population members of the variability of interest. The more the variable of interest varies, the more difficult the estimation problem. Variables related to data quantity and the problem difficulty can be assessed from the sample, but direct information for the data defect correlation, $\rho_{R,G}$ cannot be assessed from the sample. Intuitively, $\rho_{R,G}$, is the measure of the correlation between the indicator variable, which takes value 1 if the population unit is selected and 0 otherwise, and the variable of interest. If the sampling is random, with no selection bias or data defects, then $\rho_{R,G} \approx N^{-1/2}$, where N is the total population. In the Indian context, where $N \approx 1.3 \times 10^{9}$ people, this implies that the data defect correlation has to be extremely small, $\rho_{R,G} \approx 2.7 \times 10^{-5}$, for the Survey to be representative. Moreover, Meng shows that if we have an estimate of $\widehat{\rho_{R,G}}$, then we can compute the **effective sample size or Bias adjusted sample size,** n_{eff} of the Survey. In particular,

$$n_{eff} = \frac{n^{*}_{eff}}{1 + \left(n^{*}_{eff} - 1\right)N^{-1}},$$

$$where\ n^{*}_{eff} = \left(\left(\widehat{\rho_{R,G}}\right)^{2} \times \frac{1 - f + CV_{W}^{2}}{f}\right)^{-1},$$

$$N = Total\ Population,$$

$$\left(\widehat{\rho_{R,G}}\right)^{2} = \frac{(\overline{G_n} - \overline{G_N})^{2}}{\left(\frac{1 - f + CV_{W}^{2}}{f} \times \sigma_{G}^{2}\right)},$$

$$\sigma_{G}^{2} = \overline{G_N} \times (1 - \overline{G_N})\ \ for\ proportions.$$

$$Reduction\ in\ sample\ size\ or\ estimation\ efficiency = \frac{Sample\ Size - Bias\ Adjusted\ Sample\ Size\ (n_{eff})}{Sample\ Size} \times 100.$$

Intuitively, the bias-adjusted sample size is the size of the simple random sample that would have generated the same mean squared error. The expected value of the $\left(\widehat{\rho_{R,G}}\right)^{2}$, which is estimated unbiasedly as $\left(\widehat{\rho_{R,G}}\right)^{2}$, "represents the amount of selection bias (squared) expected on average from a particular recording mechanism and a chosen weighting scheme."[10]

Example to compute the data defect correlation and the corresponding bias-adjusted sample.

Take, for example, the weighted estimate of the proportion of the rural population in the NSS 68th round for the Employment and unemployment survey (see **Table 2**). The TSE, which is the difference between the Survey estimate and the value from the Census, $\overline{G_n} - \overline{G_N} \approx 0.0231\ or\ 2.31\%$. We decompose this difference into three components,

$$\underbrace{\overline{G_n} - \overline{G_N}}_{Total\ survey\ error} = \underbrace{\rho_{R,G}}_{Data\ Quality} \times \underbrace{\sqrt{\frac{1-f+CV_W^2}{f}}}_{Data\ Quantity} \times \underbrace{\sigma_G}_{Problem\ Difficulty},$$

where f is the fraction of the population sampled, which equals $\frac{456999}{1210854977} \approx .00038\ or\ .04\%$. From the sample weights, we estimated the $CV^2 \approx 2.683$. This implies that DO, which is the square of the data quantity term and is $\approx$ 9758. Based on the census proportion of the rural population of 0.6886, the $\sigma_G^2 = (0.6886) \times (1 - 0.6886) \approx$ 0.2144. Once we plug these values into the above equation, we can compute the $\widehat{\rho_{R,G}} \approx 0.0005 > 2.7 \times 10^{-5}$. We can now use these values to get the bias-adjusted sample size, n_{eff}:

$$n_{eff} = \frac{n_{eff}^*}{1 + \left(n_{eff}^* - 1\right)N^{-1}},$$

$$where\ n_{eff}^* = \left(\left(\widehat{\rho_{R,G}}\right)^2 \times \frac{1-f+CV_W^2}{f}\right)^{-1},$$

once we insert these values, we get $n_{eff}^* = ((0.0005)^2 \times 9758)^{-1} \approx 403$, and the denominator, $1 + \left(n_{eff}^* - 1\right)N^{-1} \approx 1$, where N = 1,210,854,977. We compute the bias-adjusted sample size from these values, $n_{eff} \approx 403$.

Based on the available information for the coefficient of variation of the sample weights, we can also approximate the effective sample size following Kish (1965)[11,2], defined as $n_w = \frac{sample\ size\ (n)}{1+CV^2} = \frac{456999}{1+2.683} \approx 124083$.

3. Data

The data set for our analysis comes from two nationally representative National Samples Surveys (NSS): The NSS 68th round for Household Consumption Expenditure, which surveyed 464960 individuals, 285976 (61.5%) in rural areas and 179164 (38.5%) in urban areas, and the Employment & Unemployment survey, which surveyed 456999 individuals, 280763 (61.4%) in rural areas and 176236 (38.6%) in urban areas. Both these surveys were conducted separately between July 2011–June 2012. The Ministry of Statistics and Programme Implementation (MOSPI), Government of India, had undertaken these surveys. The sampling design, shared across all the surveys, was a stratified two-stage sampling, where the sample frame for selecting the primary sampling unit (PSU) was based on the 2001 census. The PSUs were villages in rural areas and Census Enumeration Blocks in urban areas. The NSS provides the sampling weights for each household in the sample, which is used for population-level estimates. Therefore, based on the sampling frame, weights and given that unit-level data is available for each Survey, the proportion of the rural population, the proportion of the Scheduled Caste (SC) population, and the proportion of the working-age (age between 15 and 59 years) population. We also separately analyze the proportions of SC and working-age populations in rural and urban areas. This helps us to quantify the data quality and the representativeness of the NSS for the type of residence. It is essential to mention that all our estimates in the analysis are weighted estimates. For population estimates, we use values from the Census 2011[9].

4. Results

Our first set of results relates to the proportion of the Rural population. Our results suggest that the NSS has a rural bias in terms of representation. For the NSS Household Consumption Expenditure, the estimate of the rural population from the Survey was 71.43%, while the value from the Census 2011 was 68.86%. This seemingly small difference of 2.57%, based on the formulation by Meng (2018)[3], would have been generated by a simple random sample of size 324, which is in sharp contrast to the actual sample size of 464960, suggesting a reduction in the sample size or equivalently the statistical efficiency by 99.9% (see **Table 1**). Similarly, for the NSS related to the employment and unemployment survey, we find that difference between the estimate from the Survey and the value of the Census for the proportion of the rural population was 2.31% suggesting a reduction in the sample size from 456999 to the bias-adjusted simple random sample of size 403 that would have generated the same square of the difference, a decrease of 99.9% in the statistical efficiency of the estimate (see **Table 2**).

In addition to the proportion of the rural population, we compared the value of the Census with the estimates of the Survey for the SC population. We found that the NSS household consumption expenditure, and the employment & unemployment, overestimated the SC population by 2.40% and 2.22%, respectively, suggesting a reduction in a bias-adjusted simple random sample of size of 240 and 281. These results imply a reduction of statistical efficiency by 99.9% (see **Tables 1 & 2**).

In addition, when we compare the estimates of the Survey with the value of the Census for the proportion of the working-age (age between 15 and 59 years) population, we find that both surveys overestimate the working-age population, suggesting a reduction in the statistical efficiency by more than 99.8%.

However, to test if the NSS was representative at the rural and urban levels, we compared the estimates from the Survey of the proportion of the SC population and the working-age population with values from the Census separately for rural and urban areas. For rural areas, we found that both the proportion of the SC population and the working-age population are overrepresented in both the NSS, which would warrant a reduction in the statistical efficiency by more than 99.8%. However, for urban areas, both surveys overestimated the SC population; 2.01% for household consumption expenditure and 1.74% for the employment & unemployment survey, suggesting a reduction in statistical efficiency by 99.8%. In addition, both surveys overestimate the proportion of the working-age population in urban areas. However, the difference is higher for the consumption survey than the employment & unemployment survey, at 1.39% and 0.71%, respectively. The difference is, nevertheless, significant enough to warrant a reduction in statistical efficiency of approximately 99.3% for the consumption expenditure survey and 97.4% for the employment & unemployment survey (see **Tables 1 & 2**).

5. Discussion

For a survey to be representative, the data defect correlation between the indicator variable that takes the value 1 if the population unit is sampled and 0 otherwise, and the variable of interest should proximately be of the order $N^{-1/2}$, where N is the total target population. In the Indian context, the total population is $N \approx 1.3 \times 10^9$ people, which implies that the data defect correlation for the survey to be nationally representative must be extremely small, $\approx 2.7 \times 10^{-5}$. In this paper, we have quantitatively examined the data quality of the NSS by estimating the data defect correlation. We find evidence of data defect correlation to be significantly higher than what it should be $\approx 2.7 \times 10^{-5}$, for certain variables of interest (see **the method section** for an example of the calculation of the data defect correlation). This raises doubt concerning the representativeness of the NSS. We found evidence that the NSS overrepresents the rural population, the Scheduled Caste (SC) population, and the working-age (age between 15 and 59 years) population compared to the Census done during the same period. Quantitatively, our data quality analysis suggests a reduction in the statistical efficiency ranging from 97% to 99.9%, implying that the NSS is statistically non-representative at the national level, including at the rural and urban levels. Broadly this suggests that the estimates of poverty from the NSS household consumption expenditure might be exaggerated, given that it overrepresents the rural and the socially disadvantaged, the Scheduled Caste. The NSS estimates from the employment & unemployment for people engaged in the agricultural sector might be exaggerated as it overrepresents the rural population, as agriculture is predominantly a rural activity. Moreover, the NSS employment & unemployment surveys might exaggerate the workforce participation rate because it overestimates the proportion of the working-age population at the national level and when considering rural and urban areas separately.

It is important to mention that we are not the first to discuss the rural bias in the NSS estimates. In the context of poverty estimates, the rural bias of the NSS of household consumption expenditure was observed by Sundaram and Tendulkar [2003][12] and Sundaram [2013][13]. Their solution was to estimate poverty separately for rural and urban areas and use the projection from the Census estimates to estimate poverty at the national level. This correction to the sampling weights assumes that the population projection estimates from the previous censuses are reliable. Unfortunately, this is incorrect, as Sundaram [2013][13] himself documents that the proportion of the urban population based on a projection from the Census 2001 grossly underestimated the pace of urbanization. The proportion of the urban population that was projected to be achieved by 2016 was already achieved by 2011, according to Census 2011. Moreover, our analysis suggests that estimates from the rural and urban samples of the Survey itself are not representative.

Our analysis also has implications for survey strategy in general, with consequences for surveys that use the same sampling strategy, such as the National Family Health Survey (NFHS) in 2019–21[14] and the Periodic Labour Force Survey (PLFS) in 2021–22[15]. Both these Surveys use Census 2011 for the sampling frame. Suppose the urbanization process was as rapid as in the previous decade (which is very likely true). In that case, our analysis suggests that both these surveys will have a rural bias because the sampling frame does not account for the dynamic changes in the target population. Therefore, the estimates from the Survey might not be representative.

Given the importance of the Census estimate as the benchmark to estimate the data defect, it is imperative to talk about the data quality of the Census. Since 1951, the Census organization in India has evaluated the data quality of the population Census by conducting Post Enumeration Surveys (PES), also referred to as Post Enumeration Checks (PEC), after every round of the Census[16]. The primary objective of the PES is to quantify the level of omission and or double counting in the Census enumeration, as well as to quantify the response errors for certain select variables that are collected from the individuals, such as gender, age, literacy, characteristics of workers and non-workers etc. From the perspective of this paper, we will focus on the level of omission and double counting, and in particular, the difference between them, which is the net omission rate (NOR). The PES revealed that overall, the Census in 2011 had an undercount of 23.08 persons per 1000 persons and a double count of 0.10 persons per 1000 persons. Therefore, the NOR was approximately 23 persons per 1000 persons. However, the PES also revealed that the NOR was much higher in urban areas at 29 persons per 1000 persons compared to rural areas, where it was 20 persons per 1000 persons. Moreover, the survey also revealed that the NOR in heavily urbanized regions such as Delhi was approximately 57 persons per 1000 persons. The key implication from the information based on the PES is that the proportion of the rural population is much higher in the Census than what it might be. Therefore, our results which relate to the difference in the estimates of the proportion of the rural population from the NSS and the Census, potentially underestimates the true difference. Therefore, based on the information available, the data defect correlation could be much higher than what we have estimated.

The PES also revealed another relevant finding concerning age. The NOR was the highest for the younger age group of 0 to 4 years, 33 per 1000 persons and the lowest for the age group 50 to 54 years and 55 to 59 years at 13 per 1000 persons. An implication from the PES was that the Census 2011 overestimated the proportion of people in the working age group 15 to 59 years, which was true in rural and urban areas. Our estimates, therefore, of the difference between the estimates of the proportion of the working-age population in the NSS and the Census 2011 is a lower bound. The true difference is likely much higher based on available information and assessment. In summary, the key implication from the PES of Census 2011 was that the data quality of the NSS is perhaps worse than what we have estimated.

Recent developments in India emphasizing digitization suggest a movement towards data-driven policymaking.[17] Given the importance of data in framing policies, adequate attention must be paid to data quality. Otherwise, there is a possibility of misguided policies that are based on biased estimates, which might not reflect society's true changes or progress.

6. Limitations

A key limitation of the paper is that we could only analyze two surveys conducted the closest to the Census undertaken in 2011. Moreover, it would have been more appropriate had we had direct measures of poverty from the Census as from the NSS. Given that Census is such an exhaustive exercise of more than 1.3 billion people, it is practically impossible to do a detailed household consumption expenditure as in the NSS. As a result, we limited our analysis to variables, the proportion of the rural population, the proportion of the SC population and the proportion of the working-age population consistent across the NSS and the Census. However, research on poverty and employment has shown that these variables correlate with poverty, the proportion of people engaged in agricultural activities and labour force participation rate estimates, etc.[13]

Another limitation of the paper is that we cannot decompose the nature of the data defect of the NSS into dynamic changes in the target population or the selection bias resulting from differences in response rates across different sections of society. For example, there is a possibility that the SC population could have a different response rate than the general category; people residing in rural areas might be more responsive than those in urban areas. However, we believe that if the survey estimates have to become more representative and robust, we would need further rigorous research to understand the nature of data defects.

In this paper, we have limited ourselves to analysis at the national level and different types of residence (rural and urban). However, a similar analysis can be done at the sub-national level, for example, the various states, given that government policies vary across states. We have another paper in progress that looks at the state-level analysis.

Conclusion

In this paper, we have highlighted data quality issues related to the NSS that policymakers and academics have extensively used to estimate poverty and levels of unemployment. Unfortunately, there is a great deal of emphasis on data quantity and increasing the frequency of the data, with minimal discussion of data quality. There is practically a consensus among academics and policymakers on the robustness and representativeness of the NSS survey methodology, which is replicated across other surveys such as the NFHS and the PLFS. Through this paper, we wish to highlight the importance of data quality and remind policymakers and academics that the bigness of the data cannot address issues related to data quality. On the contrary, it makes us "precisely wrong".[3]

References

1. Ministry of Statistics and Program Implementation | Government Of India. https://www.mospi.gov.in/.
2. Meng, X.-L. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **12**, (2018).
3. Meng, X.-L. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **12**, 685–726 (2018).
4. Bradley, V. C. *et al.* Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* **600**, 695–700 (2021).
5. Kreuter, F. What surveys really say. *Nature* **600**, 614–615 (2021).
6. Boyd, R. J., Powney, G. D. & Pescott, O. L. We need to talk about nonprobability samples. *Trends Ecol. Evol.* **38**, 521–531 (2023).
7. India - Household Consumer Expenditure, NSS 68th Round Sch1.0 Type 2 : July 2011 - June 2012, Type - 2. http://microdata.gov.in/nada43/index.php/catalog/126.
8. India - Employment and Unemployment, July 2011- June 2012, NSS 68th Round - Overview. http://microdata.gov.in/nada43/index.php/catalog/127/study-description.
9. Home | Government of India. https://censusindia.gov.in/census.website/.
10. Unrepresentative big surveys significantly overestimated US vaccine uptake | Nature. https://www.nature.com/articles/s41586-021-04198-4?msclkid=ede012a7cfb711ec8629d39d2093074d.
11. Kish, L. Survey Sampling. *Biom. Z.* **10**, 88–89 (1968).
12. NAS-NSS Estimates of Private Consumption for Poverty Estimation: A Further Comparative Examination on JSTOR. https://www.jstor.org/stable/4413137.
13. Some Recent Trends in Population, Employment and Poverty in India: An Analysis on JSTOR. https://www.jstor.org/stable/24583397.
14. National Family Health Survey. http://rchiips.org/nfhs/about.shtml.
15. Annual PLFS Reports | Directorate General of Employment (DGE). https://dge.gov.in/dge/reference-publication-reports-annual.
16. India - Census of India 2011, Report on post-enumeration survey. https://censusindia.gov.in/nada/index.php/catalog/1366.
17. Press Information Bureau. https://www.pib.gov.in/PressReleasePage.aspx?PRID=1910063.

Table 1: Results NSS 68th Round Socio-Economic Survey [July 2011–June 2012]: Consumer Expenditure

	Sample Size	Census 2011 Value	Estimates from the NSS survey	Difference	Bias Adjusted Sample Size†
	(1)	(2)	(3)	(4) = (3) - (2)	(5)
Overall (Census Population: 1,21,08,54,977)					
Proportion of Rural Population	4,64,960	68.86%	71.43%	2.57%	324 (99.9%)
Proportion of SC population	4,64,960	16.63%	19.03%	2.40%	240 (99.9%)
Proportion of Working age (15 to 59 years) population	4,64,960	60.29%	62.41%	2.11%	536 (99.9%)
Rural (Census Population: 83,37,48,852)					
Proportion of SC population	2,85,796	18.45%	20.80%	2.35%	272 (99.9%)
Proportion of Working age (15 to 59 years) population	2,85,796	58.04%	60.71%	2.66%	343 (99.9%)
Urban (Census Population: 37,71,06,125)					
Proportion of SC population	1,79,164	12.60%	14.62%	2.01%	272 (99.8%)
Proportion of Working age (15 to 59 years) population	1,79,164	65.27%	66.65%	1.39%	1,180 (99.3%)

†The reduction in sample size is equivalent to reduction in estimation efficiency. The reduction of the sample size or estimation efficiency (in %) are reported in the brackets. $Reduction\ in\ sample\ size\ or\ estimation\ efficiency = \frac{Sample\ Size - Bias\ Adjusted\ Sample\ Size}{Sample\ Size} \times 100.$

***Note:** The difference between the true value and the estimate from the Survey is very small, such that the Survey's sample size for this variable is more than adequate. If the difference is 0, then the effective sample size would be the size of the population.

Table 2: **Results NSS 68th Round Socio-Economic Survey [July 2011–June 2012]: Employment and Unemployment**

	Sample Size	Census 2011 Value	Estimates from the NSS survey	Difference	Bias Adjusted Sample Size†
	(1)	(2)	(3)	(4) = (2) - (1)	(5)
Overall (Census Population: 1,21,08,54,977)					
Proportion of Rural Population	4,56,999	68.86%	71.16%	2.31%	403 (99.9%)
Proportion of SC population	4,56,999	16.63%	18.85%	2.22%	281 (99.9%)
Proportion of Working age (15 to 59 years) population	4,56,999	60.29%	61.95%	1.66%	870 (99.8%)
Rural (Census Population: 83,37,48,852)					
Proportion of SC population	2,80,763	18.45%	20.67%	2.22%	305 (99.9%)
Proportion of Working age (15 to 59 years) population	2,80,763	58.04%	60.32%	2.28%	469 (99.8%)
Urban (Census Population: 37,71,06,125)					
Proportion of SC population	1,76,236	12.60%	14.35%	1.74%	361 (99.8%)
Proportion of Working age (15 to 59 years) population	1,76,236	65.27%	65.97%	0.71%	4,548 (97.4%)

†The reduction in sample size is equivalent to reduction in estimation efficiency. The reduction of the sample size or estimation efficiency (in %) are reported in the brackets. $Reduction\ in\ sample\ size\ or\ estimation\ efficiency = \frac{Sample\ Size - Bias\ Adjusted\ Sample\ Size}{Sample\ Size} \times 100.$

***Note:** The difference between the true value and the estimate from the Survey is very small, such that the Survey's sample size for this variable is more than adequate. If the difference is 0, then the effective or bias-adjusted sample size would be the size of the population.