The quality of age data: Comparison between two recent Indian censuses 2001–2011

Akhilesh Yadav^a, Minakshi Vishwakarma^{b,*}, Shekhar Chauhan^b

^a Dept. of Community Medicine, Institute of Medical Sciences, Banaras Hindu University, Varanasi, 221005, India

^b International Institute for Population Sciences, Mumbai, 400088, India

ARTICLE INFO

Keywords:

Age heaping
Digit preference
Data quality
Census
India

1. Introduction

The census is one of the most valuable sources of information on demographic data. The census is a systematic procedure of recording information about a population in a definite area or boundary. The census facilitates the decision-making of the policymakers and thus helps common citizens of the country. The Indian census when first implemented in 1872 faced a conundrum situation, but since then it has come a long way. It is static as well as dynamic in nature. Static in the sense that it gives a broad picture of a country at a particular time point and dynamic within the meaning that every new census data can be compared to previous censuses. The dynamic nature of census helps in assessing the magnitude and direction of the demographic indicators and also it helps in evaluating implemented policies and paves the further course of action for the policymakers.

All in all, the census is of utmost important, and the usefulness of census cannot be denied. Keeping the important factor aside, there is one important aspect, which needs serious concern is the reliability of census data regarding age. Information on age data collected in census or other surveys suffers from various reporting errors from age not reporting to digit preference while recording the age. Usually, the census suffers from age under enumeration and age distortion due to people's tendency to rounding off or liking or disliking some digit, but apart from these facts census also suffers from age not reporting.

In the census, usually the head of the household provides information regarding all household members, and it is assumed that the head of the household knows the correct ages of all the members of the household. But in Indian society people do not give importance to

knowing their age, knowing other's correct age seems a doubtful condition.¹ Therefore, there is a high chance of reporting just an estimate of ages or some time not reporting the ages of other household members.² Several studies have been done on errors in age misreporting such as age preference or digit preference in the census.^{1,3–7} The age in the Indian census has been recorded in numbers, and that gave rise to one of the most common human biases known as digit preference. It is human tendency to opt for certain digits while getting their age recorded in the census. It is a grave concern and needs overhauling. Age is one of the most important demographic attributes that suffer human biases. It is necessary to evaluate data and to identify the types of age reporting errors before analyzing other things because these errors may affect the other estimates. Age distribution plays an important role in analyzing various demographic parameters like estimates of fertility, mortality, and marriage, etc. In more than a century history of the census, digit preference remained a problem by large. It was only in 2011 census, date of birth was recorded for the first time and before that only age has been registered in the census. The broad objective of this paper is to access the quality of age reporting in the census and also to find if there is any improvement in age reporting in India and states between 2001 and 2011. Errors in reporting age in the censuses can be seen in under-enumeration of children at age 0 and 1, overstatement of age at older ages to get benefits of government-run old-age benefits schemes, heaping at particular digits like 0 and 5, avoidance of certain digits like 3, and reports of persons of unknown age.

* Corresponding author.

E-mail address: minakshiips@gmail.com (M. Vishwakarma).

<https://doi.org/10.1016/j.cegh.2019.09.005>

Received 20 November 2018; Received in revised form 7 August 2019; Accepted 20 September 2019

Available online 01 October 2019

2213-3984/ © 2019 INDIACLEN. Published by Elsevier, a division of RELX India, Pvt. Ltd. All rights reserved.

2. Background of the study

Data on age are essential for understanding population dynamics. Population projections and other population estimates like fertility, mortality, or migration can be studied with the help of age data. The age data can pave the way for studying the distribution of the population in any field like estimates of the school-age population, estimates of the number of voters, estimates of people entering the labor force, and so on. Age data distribution is also very helpful in planning social services and implementing social policies. The existence of misreporting of age data is a stumbling block to proper planning and decision-making. Age misreporting is a very recurrent problem in developing societies. So, it becomes imperative to understand the extent of age misreporting in India.

3. Data and methodology

This study utilized data from last two rounds of censuses conducted by Office of the Registrar General & Census Commissioner, India. The socio-cultural series of census 2001 and 2011 for single year age group are used to full fill the objectives.

To access the quality of age data first, we calculated the rate of age not stated (per 100000 Population). Since not reporting age is a visible problem in developing countries and is a serious phenomenon. To access the age heaping, we have used two different indicators' Whipple's index and Myer's blended index. Whipple's index measures the digit preference of age ending with digit 0 and 5, while Myer's blended index gives the liking and disliking for all digit from 0 to 9.⁸ Both the indicators represent two different phenomena of digit preference, and together they provide a wider picture of age heaping.

Whipple's index measures digit preferences for the ages ending with 0 and 5 in a given population. It measures the digit preferences only between the age of 23 and 62. Outside this range, shifting and other problems tend to confuse the pattern of age heaping.

$$WI = \frac{500 \times (P_{25} + P_{30} + P_{35} + \dots + P_{60})}{(P_{23} + P_{24} + P_{25} + \dots + P_{62})}$$

The value of the index is divided into 5 categories ranging from highly accurate to very rough. If $WI > 175$ data falls in the very rough category if WI ranges between 125 and 175 it shows data is rough, if $110 < WI < 125$ then data represent approximate data, and if WI values lie between 105 and 110 it means data is accurate. As WI value decreases the quality of data seems to be good. If WI value comes under 105 that shows data is highly accurate.

Myer's blended index measures the extent of preference or dislikes for all digit 0 to 9. It takes care of the effect of rapid changes in fertility or mortality in any population. The principle employed in this method is as follows, first to count the number of people at ages ending with each of the 10 digits in the population beginning with age 10 and ending at age 89. Second to do the same thing beginning with age 20 separately, third to blend them in the ratio of $x+2$ and $10-x-1$ for the number ending with digit x . Fourth, compute the percentage of the blended sum to total, and fifth take the absolute deviation of this percentage from 10. Finally, compute the average of the results over the 10 digits. The value of the index ranges between 0 and 90. The higher the value of the absolute deviation for a digit, higher the preference or avoidance for the digit.

4. Results

Table 1 shows the first indicator to assess the quality of census data which is the rate of age not stated (per 100000) population. At the national level, age not reported is increased from 266 persons per 1 lakh population in 2001 census to 371 persons per 1 lakh population in the 2011 census. The rate of age not stated for males is 381 in 2011 census, and for the female, it is 360, which is higher in compared to the

Table 1

Rate of age not stated (per 100000 Population) by sex, in India and major, 2001 to 2011.

States	2001			2011		
	Total	Male	Female	Total	Male	Female
Jammu & Kashmir	451	463	437	121	147	91
Himachal Pradesh	242	266	218	148	154	142
Punjab	385	403	365	150	152	147
Uttarakhand	209	227	191	165	181	147
Haryana	381	401	357	124	125	124
Rajasthan	509	500	518	393	387	400
Uttar Pradesh	443	466	419	812	829	793
Bihar	214	240	186	388	406	368
West Bengal	139	151	127	123	136	110
Jharkhand	124	150	95	354	368	339
Orissa	169	184	153	284	289	279
Chhattisgarh	135	150	119	90	94	87
Madhya Pradesh	264	274	254	114	115	113
Gujarat	93	101	84	395	399	391
Maharashtra	122	130	113	363	377	348
Andhra Pradesh	175	193	157	912	913	911
Karnataka	98	107	89	75	82	68
Goa	520	532	508	144	149	138
Kerala	83	91	76	103	108	99
Tamil Nadu	687	713	662	93	96	89
India	266	282	249	371	381	360

previous census. Among major states of India, it has increased drastically in states of Uttar Pradesh, Maharashtra, Jharkhand, Gujarat, and Orissa. While other states of India have shown a decline in this indicator. In Tamil Nadu, rate of age not stated has decreased from 687 to 93 in one decade from 2001 to 2011. In Goa, Jammu and Kashmir, Haryana, Himachal Pradesh, and Chhattisgarh there is an improvement in age reporting. The increase in the age not stated can be attributed to the fact that more number of people are now getting older as when a population gets old, it has the tendency to forget their birth dates and thus age not stated rises significantly. In states with high age not reported, the decline in fertility has been observed recently, which indicates the demographic transition where with the decline in fertility population tends to get older thus increasing the age not reported. The age not reported among males is higher than females.

Fig. 1a & b shows the map of the Whipple Index for 20 major states in India. It gives the existing geographical variation in digit preference for males and females separately, and also shows the improvement in the index value from Census 2001 to 2011. At the national level, the value of the index has reached the rough category in 2011 from the very rough category in the 2001 census for both male and female, but the data is found to be more accurate in the case of the female population as compared to the male population. The decline in the value of Whipple's index is observed in all the states from 2001 to 2011 suggests that with the introduction of the recording of birth date, there has been an improvement in the overall scenario of age reporting. Recording birth date is a new phenomenon in the Indian census, and it has helped in improving the biases occurring as a result of digit preference. The remarkable improvement has come into sight in the case of male population in Himachal Pradesh. For 2011, the quality of age data for the male population in Himachal Pradesh has improved significantly and reached to the approximate data category, which was under the very rough category in the year 2001. From very rough category in 2001, all the states except Andhra Pradesh, Jharkhand, Bihar, Uttar Pradesh, and Rajasthan have attained the rough category in 2011 for both male and female population. The quality of the age data for the above-mentioned states is still very poor. For both male and female population in Kerala, Whipple's index falls in approximate data in 2011 while it was in the rough category in 2001.

Table 2a & 2b depicts the rank order for the male and female population for 20 major Indian states based on Myer's Index during two

a **b**

Fig. 1. **a** Maps shows the digit preference by Whipple's index for Males for the census year 2001 and 2011. **b** Maps shows the digit preference by Whipple's index for Females for the census year 2001 and 2011.

different Censuses period. The Myer's Index describes the deviation of the percentage of the blended population from 10 along with each terminal digit. The result found that the value of Myer's index had significantly improved for males in India from 24.6 in 2001 to 12.8 in 2011. The value of Myer's index varies from 0 to 95; where value closer

to 0 signifies the ideal condition when there is no preference or aversion to any specific digit while reporting age. Over the two time-period from 2001 to 2011, the value of Myer's index has improved significantly for all 20 states. Tamil Nadu has improved remarkably, moving from rank 5th in 2001 with a value of 19.3 to the 1st rank with a value of 4.1 in

Table 2a
Myer's index for the male population for India and 20 Major state, 2001–2011.

2001			2011		
Rank	State	Myer's Index	Rank	State	Myer's Index
1	Kerala	9.8	1	Tamil Nadu	4.1
2	Himachal Pradesh	15.0	2	Kerala	4.3
3	Haryana	15.2	3	Punjab	5.2
4	Goa	16.8	4	Maharashtra	9.0
5	Tamil Nadu	19.3	5	Andhra Pradesh	9.1
6	Maharashtra	20.0	6	Jharkhand	10.2
7	Uttarakhand	20.3	7	Haryana	10.3
8	Gujarat	21.4	8	Uttarakhand	10.7
9	West Bengal	21.5	9	Rajasthan	10.9
10	Chhattisgarh	21.8	10	Chhattisgarh	11.7
11	Rajasthan	23.4	11	Himachal Pradesh	11.8
12	Punjab	23.8	12	Madhya Pradesh	12.1
13	Orissa	25.1	13	Jammu & Kashmir	12.8
14	Madhya Pradesh	26.3	14	Goa	13.3
15	Karnataka	26.4	15	Gujarat	13.7
16	Andhra Pradesh	26.7	16	Karnataka	14.9
17	Jammu & Kashmir	26.8	17	Uttar Pradesh	15.4
18	Jharkhand	27.3	18	Orissa	15.5
19	Uttar Pradesh	32.4	19	Bihar	17.8
20	Bihar	33.4	20	West Bengal	19.3
	India	24.6		India	12.8

Table 2b
Myer's index for the female population for India and 20 Major state, 2001–2011.

2001			2011		
Rank	State	Myer's Index	Rank	State	Myer's Index
1	Kerala	11.9	1	Kerala	5.4
2	Haryana	14.8	2	Goa	5.9
3	Himachal Pradesh	17.8	3	Himachal Pradesh	7.8
4	Goa	19.2	4	Gujarat	9.3
5	Gujarat	19.5	5	Punjab	10.7
6	Uttarakhand	20.3	6	Tamil Nadu	11.3
7	Chhattisgarh	20.9	7	Haryana	11.4
8	Rajasthan	21.4	8	Uttarakhand	11.5
9	Punjab	21.6	9	Maharashtra	11.7
10	Maharashtra	22.6	10	West Bengal	12.1
11	Tamil Nadu	23	11	Chhattisgarh	12.6
12	West Bengal	23.4	12	Madhya Pradesh	12.9
13	Madhya Pradesh	23.7	13	Jammu & Kashmir	13.6
14	Jammu & Kashmir	25.5	14	Orissa	13.8
15	Jharkhand	25.7	15	Uttar Pradesh	14.7
16	Uttar Pradesh	25.8	16	Rajasthan	14.9
17	Orissa	26.1	17	Jharkhand	15.7
18	Andhra Pradesh	26.7	18	Karnataka	16
19	Bihar	28.1	19	Andhra Pradesh	17.1
20	Karnataka	29.3	20	Bihar	17.7
	India	23.8		India	13.2

2011. Though West Bengal has improved but very poorly and as a result of that it sloped from 9th rank in 2001 to the last rank in 2011. Punjab is another state that has shown tremendous improvement in its ranking from 12th in 2001 to 3rd in 2011. The result found that the value of Myer's index has significantly improved for females also in India from 23.8 in 2001 to 13.2 in 2011. Kerala has remained on top during both the time period and has shown improvement in the value of Myer's index. Goa is another state that has improved significantly in terms of Myer's index value. Goa ranked 4th with a value of 19.2 in 2001 and reached to the rank 2nd in 2011 with a value of 5.9. Himachal Pradesh

remained 3rd during both the time-periods. The States Uttar Pradesh, Jharkhand, Karnataka, Andhra Pradesh, and Bihar are at the bottom of the rank order during both the time-periods.

Fig. 2a and 2b describe the deviation of the percentage of the blended population from 10 along with each terminal digit for the census year 2001 and 2011 respectively among males and females. The preference of digit ends with 0 is most in the case of the male and female population for both the year but the deviation from 10 has decreased for terminal digit zero from 2001 to 2011. The next preferred digit is 5 in both the census year. The preferred digit 0 and 5 indicates that most numbers of males and females prefer to report their age with the digit ending either 0 or 5. The digit 0 and 5 give a sense of round figure, and this may be a reason for such a high reporting. The least mention terminal digits are 1 followed by 9 and 7 in the year 2001. The pattern for the neglected digit has changed in the year 2011. The strong aversion has been seen for the digit ending with 7 followed by 9. The most averted digit by the males in the year 2011 is ending with 3 indicates the social beliefs attached to the digit 3. In Indian culture, many bad omens are attached with the number ending with digit 3, and this is the reason why people try to averse the digit 3.

5. Discussion

This article has shed some light on the ongoing misreporting of age in the census by capturing age not reported and age heaping by different methods. The rate of age not reported has increased from the year 2001 at the national level, and it is higher among males than females in all major states of India. The rate of age not stated has decreased in more than half of the states. Only a few states reported the increase in this indicator. Like other previous studies, our study also revealed the existence of age heaping in the Census data.⁹ Only Kerala comes under approximate data category of data quality for both male and female population according to the Whipple index; most of the states fall under the rough or very rough categories. Consistent with the previous findings, this study also identified 0 as the most preferred digit and 9 as the most avoided digit.¹⁰ The age heaping phenomenon seems to decrease in the last decade in all the major states and at the national level. Whipple's and Myer's blended index, both give a similar result.

There is a greater need to understand the causes of the increase in the age not reported. High age not reported may be attributed to the fact that only the head of the household or any available elder person in the household reports the age of all the person in the household. There are very high chances that a single individual may not be knowing the age of all the person in the household and it is highly doubtful that the information collected from him will be correct. In Indian society, where one's age is not important, there is a high chance that he may not report age of other household members.¹ Misrepresentation of age may also arise from the ignorance of the correct age or people may tend to either understate or exaggerate their age after attaining a certain age for taking advantage of social policies. The first time in the recent Indian census (2011), date of birth was recorded, before that only age was used to record and this turned into the visible difference in the data quality. Digit preference among the population has decreased over the period of 10 years. Not only the census suffers from the age heaping, but other demographic surveys in the country also face the digit preference.^{11–14} Few studies also try to explore the factors associated with age data quality in census and surveys. Literacy, household size, uses of calendar and interaction with the administration are found the important determinants of age heaping.^{15–17} Researchers around the world have compiled age data from national censuses, and concluded that age data is often subject to several limitations^{10 18–20}

Age is one of the most important variables that need serious consideration in all the demographic surveys and census. Errors in age reporting in census data can appear in many ways including under-enumeration of children of age 0 and 1, misstatement of age for taking advantage of government schemes, heaping at particular digits, and not

Fig. 2. a Myer's Index by terminal digit for male and female, census 2001. b: Myer's Index by terminal digit for male and female, census 2011.

reporting of their age. The quality of age data needs serious intervention as age distribution plays an important role while planning any schemes and programs by the government. An innovative method, inclusion of local calendar, high-quality training for the enumerator can help in increasing the quality of age data in any enumeration.

Policy Recommendations: In the demographic study of a population, age is an important attribute that helps in measuring population structure and is also helpful in growth rate forecasting. So, it becomes imperative to understand the need for correct reporting of age. Though the birth registration is compulsory in India, it is not followed seriously. There are well-set directives and guidelines for birth registration, but still, the outcome is not visible. The government shall increase awareness about birth registration, and shall make it complete necessity to register a birth at the hospital itself. The government shall look for alternate ways through which the accuracy of age data collection can be improved. Age data shall also be collected with the date of birth as this reduces error in digit preference, though this has been in practice in the census year 2011 but was not completely flawless as people did not tell their date of birth.

Strength and Limitation: like other studies, this study also has some strengths and limitations. This study measures age quality data not only by one index but includes two well-known indices. It also provides the extent of preference and aversion of each digit with the help of the graphs which gives the pattern of digit preference/avoidance. The major limitation of this study is, it does not include a separate analysis for the rural and urban areas when it is assumed that the quality of age data will not be identical in both the area. Also, age data typically suffer from distortion owing to preferences/avoidances for certain ages and digits due to social, cultural and legal habits and norms observed in a society. However, census data do not provide individual information. One can look in to this with the help of survey data.

Ethical approval

This article does not contain any studies with human participants or animal performed by any of the authors.

Declaration of competing interest

None

Acknowledgement

Authors declare no conflict of interest. We did not get any financial assistance to carry out this study and results are free from influence of any organization. The views expressed in this publication are those of the authors and not necessarily those of the organization.

References

1. Ewbank DC. Age misreporting and age-selective under enumeration: sources, patterns and consequences for demographic analysis. *Committee on Population and Demography*. Washington, D.C: National Academy Press; 1981 Report No.4.
2. Unisa S, Dwivedi LK, Reshmi RS, Kumar K. Age reporting in Indian census: an insight. *Paper Presented at the 26th IUSSP International Population Conference, Morocco. 27 Sept-Oct. 2009; 2009*.
3. Balasubramanian K. Type of age reporting errors in the census data of Indonesia. *Demogr India*. 1974;3(2):287–305.
4. Chandra NK. Adjustment of age data for India's census population. *Demogr India*. 1980;9(1 2):274–285.
5. Jain SP. Census single year age returns and informant bias. *Demogr India*. 1980;9(1 2):286–296.
6. Prakasam CP. On quality of age data for population count-1981, in Indian states. *Paper Submitted to the Annual Conference of Indian Association for the Study of Population, Held at Indian Institute for Management, 24th December to 27th December, 1984, Bangalore. 1984; 1984*.
7. Saxena PC, Verma KR, Sharma KA. Errors in age reporting in India, a socio-cultural and psychological explanation. *Indian J Soc Work*. 1986;47(2):127–135.
8. United Nations. *Methods of Appraisal of Quality of Basic Data for Population Estimates*.

- Manual 2. United Nation Publication Sales; 1955 No. 56.XIII.2.
9. Pardeshi GS. Age heaping and accuracy of age data collected during a community survey in the Yavatmal district, Maharashtra. *Indian J Community Med.* 2010;35(4):391–395.
 10. Nagi M, Stockwell E, Snavley L. Digit preference and avoidance in the age statistics of some recent African censuses: some patterns and correlates. *Int Stat Rev/Rev Int Stat.* 1973;41(2):165–174. <https://doi.org/10.2307/1402833>.
 11. Mukhopadhyay BK, Majumdar PK. A multivariate statistical analysis of reporting error in age data of India. *J Soc Sci.* 2009;19:57–61.
 12. Choudhary CR. A study of quality of Single year age data in India. *Seminar Paper Submitted for the Partial Fulfilment for the Master of Population Studies.* Mumbai: International Institute for population Sciences; 2006.
 13. Singh M. Data quality in Indian demographic survey. *Paper Presented at the 2013 Population Association of America Conference New Orleans.* 2013; 2013.
 14. Singh M. Understanding digit preference in India using modified Whipple index: an analysis of 640 district in India. *Int J Curr Res.* 2017;9(1):45144–45152.
 15. Pullum Thomas W. *An Assessment of Age and Date Reporting in the DHS Surveys, 1985-2003* Calverton, Maryland: Macro International Inc; 2006 Methodological Reports No. 5.
 16. Crayen D, Baten J, Paper. Global trends in numeracy: a first glance at age heaping evidence from 1820-1940. *Paper Presented at the 7th Conference of the European Historical Economics Society.* Lund, Sweden. 29th June-1 July. 2007; 2007.
 17. Becker S, Diop-Sidibé N. Does use of the calendar in surveys reduce heaping? *Stud Fam Plan.* 2003;34:127–132.
 18. Susuman AS, Hamisi FH, Lougue S, et al. An assessment of the age reporting in Tanzania population census 2012. *J Soc Sci Res.* 2015;8(1):1553–1563.
 19. Spoorenberg T. Assessing the quality of age reporting at a time of general data quality improvement: going beyond the original Whipple's index. *XXVI IUSSP International Population Conference, Morocco 27 September, 2009, Session P-5. Morocco.* 2009; 2009.
 20. Suong Y. Quality of age data by sex in censuses of some selected asian countries. *Seminar Paper Submitted as a Part of Requirements for Diploma Course in Population Studies.* Mumbai: International Institute for population Sciences; 1995.