Journal of Population and Social Studies, Volume 23 Number 1 January 2015 : 47 - 56
DOI 10.14456/jpss.2015.1

Influence of Literacy on India's Tendency for Age Misreporting: Evidence from Census 2011

Gopal Agrawal[1] and Puneet Khanduja[2]

The quality of age reporting in the recently released single-year age data from the Indian Census 2011 is examined. Besides analyzing whether there was any significant improvement in quality from 2001 to 2011, the paper investigates whether there is a relationship between growth in the literacy rate and the quality of age-reporting. Modified digit-specific and total Whipple's indices are used to check patterns in digit preferences/avoidances among Indians in the two censuses. Correlation coefficients are estimated to analyze the influence of literacy on the tendency for age-misreporting among the Indian population. The total Whipple's modified index declined from 5.5 to 2.9 between 2001 and 2011.The correlation coefficient of the association between growth in the literacy rate and quality of age reporting is significant ($r = -0.92$; $p<0.01$). We conclude that India has made remarkable progress in improving the quality of age reporting in the population Census during the last decade, and that education played a vital role. It may be expected that increased literacy will further improve the quality of age data in states and areas still lagging behind.

Keywords: *data quality; age misreporting; census; literacy; India*

Introduction

The national population census is the most detailed information source for the general population at the level of small localities, cities, districts, state and the country as a whole. For the world, India has set an example by having an unbroken series of decennial censuses since 1872, in spite of adversities such as wars, epidemics, natural disasters, political disturbances and many others. Since its inception in 1872, the Indian Population Census is the most credible source of information on demographic characteristics, economic activity, literacy and education, housing and household amenities, urbanization, fertility and mortality, ethnicity (scheduled castes and scheduled tribes), language, religion, migration, disability and many other socio-cultural and demographic indicators. Census 2011 is the fifteenth Census in this continuous series from 1872 and the seventh since Independence. In a country like India with a multiethnic, multilingual, and multicultural population, the Census is much more than a mere head count. Following the slogan "Our Census - Our Future", the Census is the basis for reviewing the country's progress in the past decades, monitoring ongoing government programs and most importantly, to plan for the future (Registrar General of India, 2011).

However, a disturbing drawback of the Census in the past has been the large differences in the quality of age reporting. Recording age is an integral part of all survey and census efforts,

[1] Department of Development Studies, International Institute for Population Sciences, Mumbai, India
Email: gopalphd.iips@gmail.com

[2] School of Public Health, Post Graduate Institute of Medical Education & Research, Chandigarh, India

Published online: 26 January 2015

and since biblical times, age determinism in demographic and epidemiological studies is well recognized. Age data has significance since most demographic, epidemiological analyses and analytical studies are performed according to age and sex variables (Borkotoky & Unisa, 2014; Pardeshi, 2010). Surveys involve many sources of sampling and non-sampling errors, of which age-misreporting is the most fundamental. In many populations with a low level of literacy, most people are not aware of their exact age or of the ages of other family members. During enumeration, such people often make guesses for their age when asked by the interviewer. In most cases, there is also a tendency to report certain preferred ages, often a number ending with certain digits (most frequently 0, 2 or 5) (Pathak & Ram, 1998).

Compared to developed countries, the incidence of inaccurate age reporting is greater in census or sample survey data from developing countries. India, Morocco and Switzerland present large differences in the quality of age reporting in their population censuses (Spoorenberg, 2009, Talib, Ali, Hamid, & Zin, 2010). Age misstatements affect various demographic and socio-economic indicators, including calculations of the age structure of the population. Apart from this, inaccurate age reporting affects the sampling strategy for surveys conducted in India such as the National Family Health Survey (NFHS), the District Level Reproductive and Child Health Household Survey (DLHS), the National Sample Survey (NSS), and the World Health Survey (WHS). It also means that inaccuracies are entered into the Sample Registration System (SRS). This is especially troubling since these data sources provide information at the lowest possible aggregation point, that is, the village and town level.

Given the significance of Census age data, a number of studies have investigated errors in age reporting such as digit preference and age preference, and have produced methods for smoothing age data. These studies strongly recommend evaluating the quality of age data before using it for analysis and planning purposes (Balasubramanian, 1974; Chandra, 1980; Ewbank, 1981; Jain, 1980; Prakasam, 1984, Zaki & Zaki 1983, Saxena, Verma & Sharma, 1986). Previous studies have documented that illiteracy has been primarily responsible for the inaccurate age reporting in the Indian Censuses. They document a number of problems with age data arising out of illiteracy such as ignorance of age, negligence in reckoning the precise age, deliberate misstatement and misunderstanding of the questions (Ambanavar & Visaria, 1975; Mukhopadhyay, 1983). In view of the fact that the mind of an educated person is trained in numeracy, and that he/she is more likely to appreciate the importance of the census, the educational level of the informant is likely to affect the quality of age data. For these reasons, it is assumed that the quality of age data will improve with the increase in literacy levels among the Indian population. Surprisingly however, the quality of age returns in the Indian censuses of the period 1951-71 deteriorated, in spite of the rapid growth of literacy and education (Ambanavar & Visaria, 1975; Mukhopadhyay, 1983; Unisa, Dwivedi, Reshmi, & Kumar, 2009).

Recently, the Registrar General of India has released single-year age data for Census 2011. Against the background described above, this paper evaluates the quality of this data and analyzes whether there is any significant improvement in the quality of age reporting over the 2001 to 2011 period. Considering that the effective literacy rate in the Indian population has increased from 65% to 74% during the same period, we also examine the association between growth in literacy and the extent of age misreporting in Indian Census 2011.

Data and Methods

This paper uses data from two successive Indian censuses undertaken in 2001 and 2011. The first assessment uses single-year age data, after which we progress to the examination of five-year age distributions (Moultrie et al., 2013). The quality of age data can be measured by means of age heaping indices, developed to detect the extent of preferences or avoidances for certain ages. There are several standard indices available for quality assessment of single - year age data such as Bachi's, Myer's, Zelnik's, and Whipple's index. However, Whipple's index is the simplest and most widely used age heaping index.

Initially, Whipple's index was developed to measure the extent of preference for ages ending with digits 0 and 5. Later, several modifications were carried out to overcome the limitation of examining only two digits. Spoorenberg and Dutreuilh (2007), and Spoorenberg (2009) well documented several modifications carried out in the Whipple's index over time, and a brief description is provided below.

The first modification was suggested by Roger, Waltisperger, and Corbille-Guitton (1981) by distinguishing between preferences for ages ending in 0 and those ending in 5. Following this, Noumbissi (1992) proposed the following equations (labeled 1 and 2) to measure age heaping for all ten digits (0 to 9):

$$\boldsymbol{W_j = 5 * \frac{\sum_{i=3}^{6} P_{ij}}{\sum_{i=0,10,20,30} {}_5^{28+i+j}P}}\ \boldsymbol{, for\ every\ j = 0, 1, 2} \qquad (\mathbf{1})$$

$$\boldsymbol{W_j = 5 * \frac{\sum_{i=2}^{5} P_{ij}}{\sum_{i=0,10,20,30} {}_5^{18+j}P}}\ \boldsymbol{, for\ every\ j = 3, 4 \ldots\ldots\ldots 9} \qquad (\mathbf{2})$$

Where,

P_{ij} is the population counts at age (ij)

${}_5^{28+i+j}P$ & ${}_5^{18+j}P$ are the total population in the age group (28+j to 32+j) and (18+j to 22+j)respectively.

W_j = 1 indicates that there no digit preferences or avoidances in the reporting of ages.

However, W_j> 1 or W_j<1 indicates a digit preferences or avoidances for digit j in question.

It should be noted that this method is not suitable for making spatial, temporal and/or other comparisons. To overcome this problem, Spoorenberg and Dutreuilh (2007) constructed the total modified Whipple's Index (W_{tot}) (equation 3 below). This is a summary index that summarizes all age preference and avoidance effects by taking the sum of the absolute differences between W_i and 1. If W_{tot} = 0, it indicates no digit preference in age data. If all persons report ages ending in the digits 0 or 5, it takes the maximum value of 16.

$$\boldsymbol{W_{tot} = \sum_{i=0}^{9} (|W_i - 1|)\ , i = 0, 1, \ldots\ldots 9} \qquad (\mathbf{3})$$

Here,
W_i = digit-specific modified Whipple's index developed by Noumbissi (1992)

Spoorenberg (2009) recommended the use of W_{tot} as the best summary measure for the assessment of the quality of age reporting and its change over time. He maintained that this summary indicator of overall age reporting quality is a highly reliable and easy to calculate measure both for comparing successive censuses in a single geographic area, such as a state or country, and for comparing the accuracy of age data from different geographic areas.

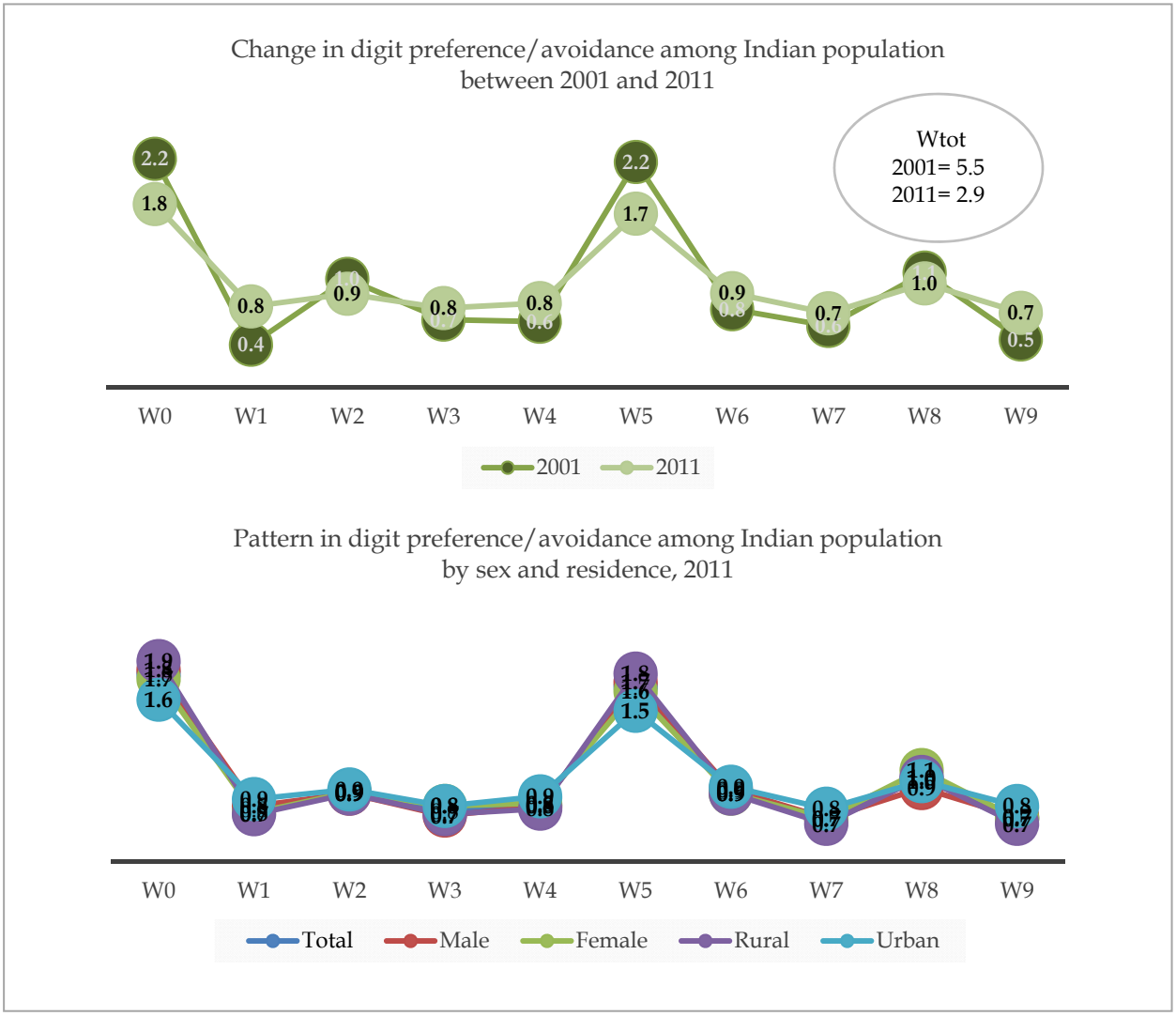
Given the merits of this index, it is used here for the comparative assessment of the Indian Census 2001 and 2011 single-year age data. Correlation analysis is then carried out to understand the relationship between the growth in literacy rates and the quality of age reporting.

Results and Discussion

Digit preferences/avoidances among Indians

Figure 1 presents the estimates of digit-specific and total modified Whipple's indices depicting the changes and patterns in digit preferences/avoidances among the Indian population by sex and residence between 2001 and 2011. Between 2001 and 2011, Whipple's modified index (W_{tot}) has declined by 47% (from 5.5 in 2001 to 2.9 in 2011), thus suggesting a significant improvement in the quality of age reporting among the Indian population. Age reporting in India follows the classic pattern of strong preference for ages ending with digits '0' and '5', as reflected by high W_i values. Strong avoidance among Indians for ages ending in 1 and 9 are reflected by W_i values less than 1. However, Figure 1 illustrates that these preferences/avoidances for certain digits weakened over the time period. The lower graph in Figure 1 depicts the patterns in digit preferences/avoidances by sex and residence in 2011; it indicates that there are similar patterns in digit preferences/avoidances for both sex and residence. Important evidence emerges here that there are large differences in the quality of age reporting between rural and urban India: people in urban India have less extreme preferences/avoidances for ages ending with certain digits than the rural population. Contrary to this, the difference between sexes is not very large; the quality of age reporting is only marginally better among men compared to women. These findings are plausible since urban people are more likely to be better educated, and hence to be well aware of their exact age and more likely to understand the importance of reporting their age accurately.

Figure 1: Quality of age reporting in Indian Census 2001 & 2011: digit-specific modified Whipple's indices (W_i) and modified total Whipple's indices (W_{tot}) estimates



Variations in Quality of Age Reporting

Table 1 presents state variations in total modified Whipple's indices (W_{tot}) by sex and residence and improvement in quality of age reporting between 2001 and 2011. The greater the value of W_{tot}, the lower the quality of age data. In 2011, the quality of age data was substantially greater in Kerala (0.9) followed by Himachal Pradesh (1.3), Gujarat (1.9), Tamilnadu (2.2), and Punjab and Maharashtra (2.3).

Table 1: Total modified Whipple's Indices (W_{tot}) by sex and residence in major states of India in 2001 and 2011

States	Total		Male		Female		Rural		Urban	
	2001	2011	2001	2011	2001	2011	2001	2011	2001	2011
Jammu & Kashmir	5.97	2.83	6.21	2.70	5.80	3.03	6.32	3.08	5.09	2.25
Himachal Pradesh	3.82	1.25	3.57	1.03	4.06	1.50	3.91	1.33	3.13	0.84
Punjab	5.34	2.26	5.61	2.43	5.00	2.17	5.47	2.53	5.05	1.84

Uttaranchal	4.94	2.46	5.01	2.42	4.87	2.52	4.98	2.61	4.86	2.13
Haryana	3.87	2.45	4.04	2.46	3.70	2.46	3.76	2.75	4.14	1.98
Rajasthan	5.28	3.40	5.68	3.70	4.85	3.18	5.25	3.67	5.34	2.76
Uttar Pradesh	6.76	3.70	7.83	4.42	5.89	3.19	6.85	3.89	6.47	3.14
Bihar	7.10	4.37	8.12	4.91	6.30	4.02	7.15	4.41	6.83	4.01
Assam	6.11	2.92	6.00	2.74	6.21	3.12	6.29	3.10	5.17	2.08
West Bengal	5.39	2.52	5.33	2.33	5.48	2.73	5.49	2.67	5.15	2.25
Jharkhand	6.16	3.66	6.43	3.74	5.88	3.56	6.34	3.95	5.69	2.84
Odisha	5.78	2.80	5.76	2.54	5.86	3.05	5.97	2.91	4.89	2.24
Chhattisgarh	4.92	2.71	5.18	2.78	4.67	2.72	4.92	2.91	4.92	2.22
Madhya Pradesh	5.75	2.96	6.22	3.23	5.23	2.72	5.83	3.17	5.52	2.40
Gujarat	4.95	1.94	5.34	2.08	4.50	1.86	4.90	2.21	5.03	1.64
Maharashtra	4.98	2.31	4.89	2.04	5.12	2.57	5.23	2.84	4.69	1.71
Andhra Pradesh	6.25	3.66	6.23	3.50	6.30	3.80	6.40	4.21	5.84	2.60
Karnataka	6.38	3.24	6.12	2.98	6.66	3.51	7.04	3.99	5.19	2.15
Kerala	2.32	0.88	2.20	0.82	2.41	0.96	2.42	0.96	2.03	0.80
Tamilnadu	4.79	2.19	4.46	1.90	5.19	2.48	5.58	2.90	3.89	1.48
ALL INDIA	5.52	2.85	5.71	2.99	5.32	2.83	5.76	3.24	5.00	2.12

On the other hand, Bihar (4.4) returned the lowest quality age data in Census 2011, followed by Uttar Pradesh, Jharkhand and Andhra Pradesh (3.7). State variations in quality of age reporting were consistent by sex and residence also. Between 2001 and 2011, W_{tot} declined substantially by 47% from 5.5 to 2.9. Improvement in quality of age reporting was greater in the states that returned greater quality single year age data in 2001; this was highest in Himachal Pradesh followed by Kerala, Gujarat and Punjab.

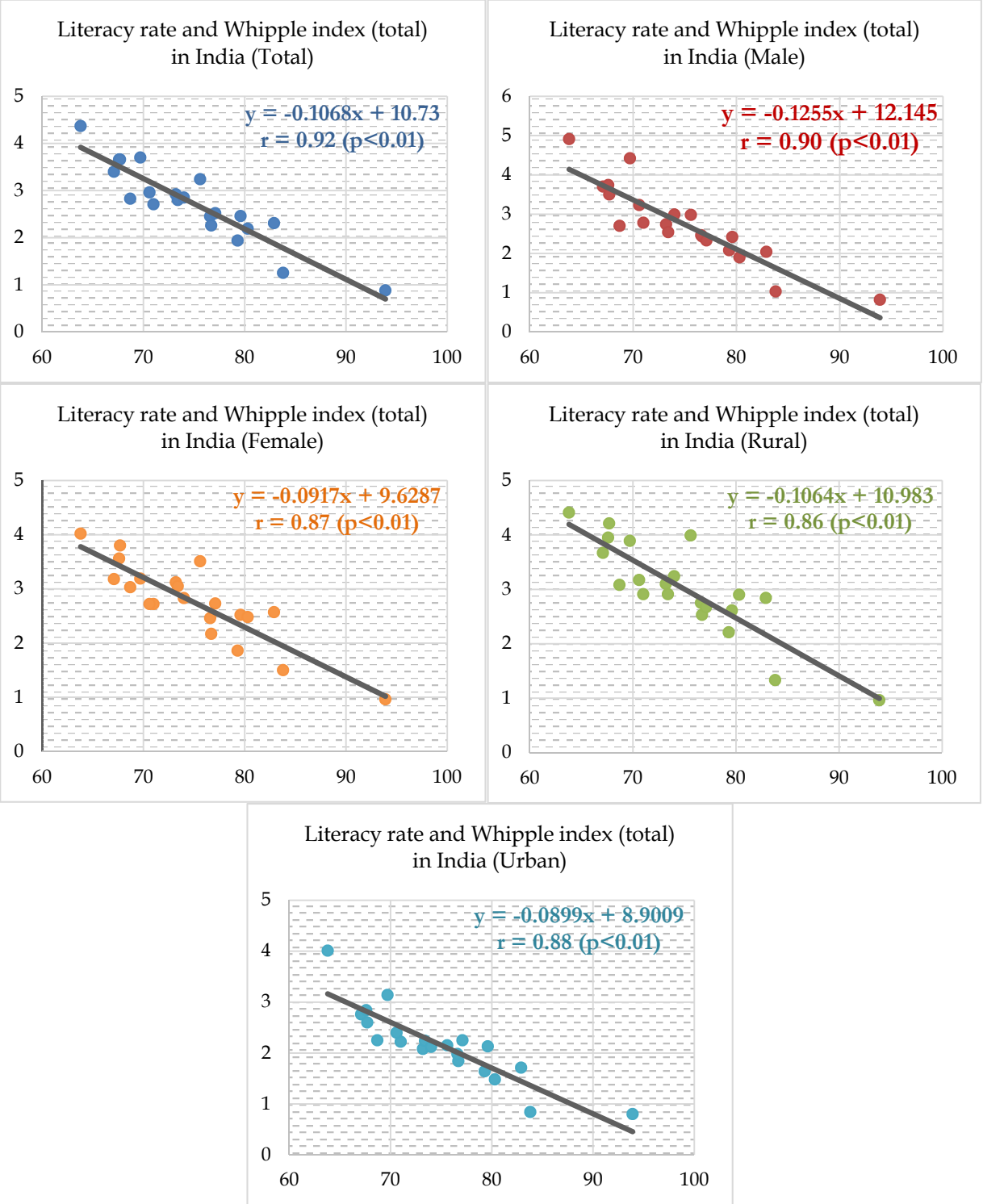
Overall in 2001, females (5.3) reported slightly better quality age data compared to males (5.7). During 2001-11, the gender gap in the quality of age reporting also declined. However, the pattern in quality of age reporting by sex varied by the level of development in states and mixed patterns were observed. In demographically advanced states (states ahead in demographic transition processes) including Kerala, Tamilnadu, Karnataka, Maharashtra, and Himachal Pradesh, women were more likely to return better quality age data compared to women in states lagging behind in these indicators. However, they still reported lower quality age data compared to their male counterparts. Contrary to this, females in some of the demographically lagging states (including of Bihar, Uttar Pradesh, Madhya Pradesh and Rajasthan) surprisingly returned better quality age data compared to males. This was a counterintuitive finding as the literacy rate was considerably higher among men in these states. There could be some confounding factors responsible for this finding, needing further exploration using micro-level information.

Results by residence depicted that urban persons were more likely to return better quality age data compared to rural persons. In 2001, the W_{tot} was marginally greater in rural India (5.7) compared to urban India (5.0). However, this gap increased during the last decade 2001-11 (W_{tot}, rural= 3.2; W_{tot}, urban= 2.1), thus suggesting that quality of age reporting improved in urban India more rapidly compared to that in rural areas. This is possibly due to higher literacy, lower negligence in reporting correct age and greater awareness of the importance of age. Further, rural-urban differentials in the quality of the Census age data were found to be roughly equivalent across all the states of India.

Correlation Analyses

Figure 2 displays the interrelationships between growth in literacy and quality of age reporting among the Indian population by sex and residence. It clearly depicts a very strong positive association between literacy and quality of age reporting among Indian states; with increasing literacy rates, quality of age reporting improved significantly. The upper left graph shows the association between extent of age-misreporting and literacy rate among the total population. As expected, it indicated that states with greater literacy tend to return better quality single-year age data. The correlation coefficient (-0.92) was statistically significant at $p<0.01$. Next, we investigate the extent to which sex and residence played a role in this association. As can be seen from the upper right graph of Figure 2, the correlation coefficient among males was -0.90 ($p<0.01$) slightly greater than that of females ($r=-0.87$, $p<0.01$). States with greater improvement in literacy among urban dwellers were likely to return greater quality age data (bottom left graph; $r=-0.88$, $p<0.01$). Moreover, the middle left graph indicates that higher growth in the rural literacy rate was also associated with improvement in the quality of age reporting (correlation of -0.86; $p=0.01$).

Figure 2: Association between growth in literacy rate and quality of age reporting by sex and residence in Indian Census 2011



Conclusion

Theoretically speaking, collection of information on age should be a very simple and easy task. Yet large differences have been observed between actual age returns in the Census and surveys and the true age for a large part of the population. While errors inevitably occur in

the collection of census data, it is essential to detect and quantify the errors by evaluation so that the users are aware of the quality of the data.

It is well established that apart from age-misreporting, age data typically suffers from distortion owing to preferences/avoidances for certain ages and digits due to social, cultural and legal habits and norms observed in a society. We found that age reporting in India followed a classic pattern, with strong preference for ages ending in the digits '0' and '5' and proportional avoidance of ages ending with digits other than these two. However, evidence shows that India must be praised for substantially improving the quality of age reporting in the Census, and that the gender gap in quality of age reporting has narrowed also.

Overall, the interrelationships between growth in literacy and changes in quality of age reporting appeared to be consistent with our conceptual background. States with greater literacy rates returned higher quality age data in Census 2011. The very interesting evidence emerging was that the association became stronger and stronger with the improvement in literacy and quality of age reporting. For instance, the total modified Whipple's index indicated that the quality of age reporting was better for Indian males compared to females; also, the association between literacy and quality of age reporting was stronger in urban India compared to rural India.

While the best possible effort was made to establish an association between growth in literacy rates and quality of age reporting in the Indian Census, the results are subject to limitations imposed by methods of data collection and estimation procedures. First and foremost, this study could not explore the adjusted effect of literacy on quality of age reporting after controlling for the effect of relevant socioeconomic and cultural determinants due to non-availability of such information. Second, the study could not explore the suitability of the total modified Whipple's index for comparison across different socio-economic strata of the population. Nevertheless, the results of the study have potential research and policy value in increasing our understanding of data quality and the potential for its improvement in the most recent Census.

References

Ambanavar, J.P.&Visaria, P. (1975). Influence of literacy and education on the quality of age returns.*Demography India*, 4 (1), 11-15.

Balasubramanian, K. (1974). Type of age reporting errors in the census data of Indonesia.*Demography India*, 3 (2),287-305.

Borkotoky, K.& Unisa, S. (2014). Indicators to examine quality of large scale survey data: An example through district level household and facility survey. *PLoS ONE*, 9(3): e90113. doi:10.1371/journal.pone.0090113.

Chandra, N. K. (1980). Adjustment of age data for India's census population.*Demography India*, 9 (1& 2), 274-285.

Ewbank, D.C. (1981).*Age misreporting and age-selective under enumeration: Sources, patterns and consequences for demographic analysis*.Committee on Population and Demography, Report No.4.Washington, DC: National Academy Press.

Jain, S. P. (1980).Census single year age returns and informant bias.*Demography India*, 9 (1&2), 286-296.

Moultrie, T., Dorrington, R., Hill, A., Hill, K., Timæus, I., & Zaba, B. (2013). *Tools for demographic estimation*. Paris: International Union for the Scientific Study of Population (IUSSP). Retrieved from http://demographicestimation.iussp.org/sites/demographicestimation.iussp.org/files/TDE_2013_2ndImpression.pdf

Mukhopadhyay, B.K. (1983). Pattern of change in age reporting during 1961-71, Indian Census data.*Demography India*, 12 (1), 131-144.

Pardeshi,G. S. (2010). Age heaping and accuracy of age data collected during a community survey in theYavatmal District, Maharashtra.*Indian Journal of Community Medicine*, 35 (3), 391-95.

Pathak, K. B.& Ram, F. (1998).*Techniques of demographic analysis*. Mumbai: Himalaya Publishing House.

Prakasam, C. P. (1984, December). *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, Bangalore.

Registrar General of India (RGI): *Census of India 2011*. Retrieved from www.censusofindia.gov.in

Roger, G., Waltisperger, D., & Corbille-Guitton, C. 1981. *Les structures par sexeetâgeenAfrique*(Structures by sex and age in Africa). Paris: Groupe de DémographieAfricaine, IDP-INED-INSEE-MINCOOPORSTOM.

Saxena, P. C., Verma, K. R.& Sharma, K. A. (1986).Errors in age reporting in India- A socio-cultural and psychological explanation.*Indian Journal of Social Work,* 47 (2), 127-135.

Spoorenberg, T.&Dutreuilh, C. (2007). Quality of age reporting: Extension and application of the modified Whipple's index.*Population (English Edition)*,62 (4), 729-741.

Spoorenberg, T. (2009, September). *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.

Talib, A. L., Ali, M. S., Hamid, M.S., & Zin, K. M. (2010).*Age reporting behaviour: A case study of 1991 and 2000 population and housing censuses*. Department of Statistics, Malaysia, 61–84.

Unisa, S., Dwivedi, L. K., Reshmi, R. S., & Kumar, K. (2009). *Age reporting in Indian census: An insight*. Paper presented at the 26th IUSSP International Population Conference, Morocco.

Zaki, P. K.& Zaki, A. J. (1983). A comparative study of age reporting in Pakistani censuses and surveys: 1951-1981.*Demography India*, 12 (1), 145-172.