NSS-12

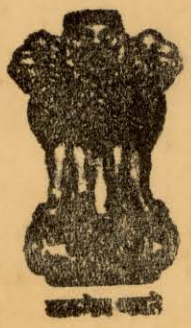
Cab. S. 15
1,000

THE NATIONAL SAMPLE SURVEY

NUMBER 12

A TECHNICAL NOTE ON AGE GROUPING

WITH A FOREWORD BY
P. C. MAHALANOBIS



Issued by
The Cabinet Secretariat : Government of India

Printed by the Eka Press, Calcutta in 1958 and published by the Manager of Publications, Civil Lines, Delhi-8

THE NATIONAL SAMPLE SURVEY

NUMBER 12

A TECHNICAL NOTE ON AGE GROUPING

WITH A FOREWORD BY
P. C. MAHALANOBIS

Issued by
The Cabinet Secretariat : Government of India

Printed by the Eka Press, Calcutta in 1958 and published
by the Manager of Publications, Civil Lines, Delhi–8

Note of Caution

Being the scanned copy of old NSS report, this document may suffer from following limitations -

i. Poor Quality of the Scanned images.
ii. Page(s) missing in between.
iii. Improper sequencing/arrangement.

ACKNOWLEDGEMENTS

The Technical Note was prepared by Ajit Das Gupta with the assistance of Samarendra Nath Mitra, and his other colleagues in the Demography Section in the Indian Statistical Institute (ISI).

The work in its final form is naturally the product of co-operative labour of men in the Statistical and Field Wings of the National Sample Survey (NSS). Specific mention may be, however, made of Pronoy Kumar Chatterjee for supervision of field work for certain special studies; Jitendra Nath Taluqdar, Gopal Chandra Bhattacharyya and Sukamal Das for supervision of computing; and Suranjan Sen Gupta for editing.

Acknowledgement is also due to persons who sent useful comments on the draft.

THE NATIONAL SAMPLE SURVEY

NUMBER 12

A TECHNICAL NOTE ON AGE GROUPING

FOREWORD

Biases in age returns occur extensively in India as in many other countries. Attention to this problem has been given in the Indian Censuses; and special groupings have been adopted from time to time for age tabulations and the smoothing of age returns. No systematic study of age distortion has, however, been made so far. It became necessary to consider this question in connexion with the analysis of the demographic data collected in the National Sample Survey (NSS). This technical note gives the results of special investigations undertaken by Ajit Das Gupta and his colleagues in the Indian Statistical Institute for a period of about three years on basis of the NSS and Census age returns, special experiments, and contemporary field studies.

2. The heaping of age returns has been studied in this report for the three components :

(1) digit preference (as such, without the effects of estimation error and age bias);

(2) estimation error (as such, without the effects of age bias); and

(3) age bias;

with a view to isolate the influence of each of these elements by itself (some amount of overlap was, however, unavoidable), and to build up the most efficient set of grouping from the knowledge so obtained.

3. The conventional 0–4 : 5–9 quinary grouping [connoted in the present note by 0 : 5] was found to be relatively inefficient for the NSS data; and the set 2 : 7 came out to be most efficient for important age-income segments of the population : this set also seemed to give a more balanced distribution of the group errors.

4. The superiority of this set was also brought out by other special investigations made by D. B. Lahiri and presented in the paper "Recent developments in the use of techniques for assessment of errors in nation-wide surveys in India" at the International Statistical Conference, Stockholm, 1957. This set had been found to be the most efficient for age returns in the Uttar Pradesh Census of 1951; in the 1931 Census Report also the 2 : 7 grouping had been recommended after an analysis of the age in individual years on traditional lines. No detailed examination could be made for the 1941 Census age data as the tabulations were based on the two per cent Y-sample. In the *Census of India 1951, Paper No. 3, 1954*, some detailed examination of the age data of Uttar Pradesh led to the 2 : 7 set being described as a "standard" grouping ; but the 3 : 8 set was recommended as "proper" for reasons not clearly understood.

5. Experience of sample surveys conducted in India suggests that with greater care and interviews at depth, which are not practicable in the Census, it is possible to make improvements in the age returns.

6. The analysis in this report was restricted to the specific objective in view. Other aspects of the quality of population data as obtained through a Census or through the NSS have not been considered here. Studies are, however, going on; and the systematic under-reporting of the population in the young age group 0–14 in the 1951 Census was, for example, examined in *NSS Draft No. 14*, "*Some characteristics of the economically active population*" on the basis of a comparison with age distributions of NSS data on population.

11 October 1958 *P. C. Mahalanobis*

THE NATIONAL SAMPLE SURVEY

NUMBER 12

A TECHNICAL NOTE ON AGE GROUPING

CONTENTS

	Page
Foreword	iii–iv
Section One : Introductory	1
Section Two : The problem	3
Section Three : Digit preference	8
Section Four : Estimation error	13
Section Five : Age bias	21
Section Six : Measures of concentration and distortion	24
Section Seven : Grouping efficiency	28
Appendix 0 : Proforma of Schedule	32
Appendix 1 : Detailed Tables	33

INDEX TO TABLES IN THE TEXT

SECTION TWO

Table 2.1 : Distribution of individuals by type of available evidence about age 5

SECTION THREE

Table 3.1 : Frequency distribution of the central missing digit supplied by guess.. .. 8

Table 3.2 : Frequency distribution of digits supplied by guess in the first two consecutive missing digit places 9

Table 3.3 : Frequency distribution of selected paired consecutive digits supplied by guess 9

Table 3.4 : Frequency distribution of all the three consecutive missing digits supplied by guess 10

Table 3.5 : Population returned at repeated digit individual ages in Census and expected population on elimination of second order of digit preference 12

SECTION FOUR

Table 4.1 : Distribution of (1) the second place after decimal of the eye-estimated length of lines and (2) the end-digit of age of all-India rural sample population aged 40-above 13

PAGE

TABLE 4.2 : Distribution of eye-estimate of the aggregate lengths of a cluster of 5 lines (actual aggregate 6.33L) rounded to the first decimal place 14

TABLE 4.3 : Distribution of individuals in age-assessed minus age-stated classes under education standard breakdowns 14

TABLE 4.4 : Distribution of individuals in age-assessed minus age-stated classes under rating of statement categories 15

TABLE 4.5 : Concentration at end-digit '0' in age statements under different rating of statement categories 16

TABLE 4.6 : Distribution of individuals in different age ranges under age-assessed minus age-stated groups 16

TABLE 4.7 : Distribution of individuals in assessment-evidence type categories under sex breakdowns 17

TABLE 4.8 : Concentration at end-digit '0' in age-assessed series under different rating of assessment classes 18

TABLE 4.9 : Frequency distribution of the number in different age groups by adjusted difference in ages 19

SECTION FIVE

TABLE 5.1 : Ratio of numbers returned at each end-digit to total numbers in the successive decennial age ranges 22

TABLE 5.2 : First differences of the ratios of numbers returned at each end-digit as shown in Table 5.1 23

SECTION SIX

TABLE 6.1 : Measures of concentration at individual end-digits and index of aggregate distortion in age returns 25

TABLE 6.2 : Relative range measures of deviation in decennial age ranges 27

SECTION SEVEN

TABLE 7.1 : Group efficiency index of different sets of grouping 28

TABLE 7.2 : Comparative deviations between Census numbers returned and expected under different sets of grouping 31

THE NATIONAL SAMPLE SURVEY

NUMBER 12

A TECHNICAL NOTE ON AGE GROUPING

*This Report, A Technical Note on Age Grouping, was prepared by the Indian Statistical Institute and is being published in the form in which it was submitted to the Government of India. The views contained in it are not necessarily those of the Government of India.**

SECTION ONE

INTRODUCTORY

1.1. The question of comparative efficiency of different sets of age grouping arose in analysis of National Sample Survey (NSS) demographic data. The feature of heaping up at certain digits, ascribed to 'digit preference', and the resulting distortion of age returns were studied at some depth in this context. The examination of the constituent data itself is no doubt of primary importance in deciding on an efficient set of age grouping, but it was felt that the precise nature of the complex of factors underlying the age distortions had to be understood clearly before the question could be properly tackled. Advantage was, therefore, taken of some experiments contemporarily organised to investigate the interplay of these factors.

1.2. The results of the investigation and the conclusions arrived at are set down in the following sections. The conventional 0–4 : 5–9 grouping, symbolised in the present note as 0 : 5, was found relatively inefficient for the NSS medium and the most efficient set 2 : 7 was adopted in grouping ages for the purpose of internal analysis and also for the purpose of presentation. The departure from the convention itself seemed to call for sufficient justification ; this note was prepared to provide the necessary logical foundation. The findings are of course of wider implication.

1.3. *Summary findings* : The digit preference was examined in isolation from other estimation errors in the Estimation and Extra Sensory Perception (E & ESP) Study 1954 and the examination extended to actual age data. The digit preference as such was seen to have little effect on age record, the estimation errors being by far the dominant factor in the Indian situation where ages were mostly recorded from guess. West Bengal Special Demography (WBSD) Study 1954 for example disclosed that definite evidence of age, including a definite statement of the date of birth and that of children, was available only for one out of six persons, while for about half the population the ages were recorded just from guess.

* The draft report (Number **D. 16**) was submitted to the Government of India in December 1956.

1.4. The digit preference comprised primarily in a tendency to keep to the middle of the digit array 0, 1, ..., 8, 9 : a liking for a run of consecutive digits and a dislike to repeat digits, for convenience called the second order of digit preference, were also found. The digit preference might be significant in situations where no major distorting factors entered.

1.5. Estimation errors on the other hand produced the familiar pattern of rounding up at digits '0' and '5'. The analysis of the estimation error suggested that apart from the errors of rounding up, there could be a bias to over-estimate. In WBSD Study, both the ages as stated by the informant independently and as assessed by the investigator from the evidence available on his best efforts, were recorded, along with the type of evidence available, the rating of statement and the rating of assessment. The age-assessed was identical with the age-stated in about 3 out of 4 cases; but for the remaining, age-assessed was higher than age-stated nearly twice or thrice as often, more often in the middle age range. The age-assessed series however did not appear to be of any better quality than the age-stated series and over-estimation in assessment was indicated. Due to general ignorance of age, the age assessed by the investigator is usually recorded and the Census age data also supported the finding. The bias to over-estimate the age was confirmed in the West Bengal Household Comparative (WBHC) Study 1955, where the ages of the common population of NSS 4th round and the Study were recorded after a lapse of three years. Significant over-estimation in recorded ages appeared in WBHC Study; the bias actually started as one of under-estimation in the young age range which changed to progressive over-estimation with increasing age, resulting in overestimation in the aggregate.

1.6. The third basic element distorting age returns, the age bias, involving conscious mis-statement of age, was difficult to locate from internal analysis alone, particularly in situations like India where estimation errors are much larger in dimension.

1.7. A modified simple measure of concentration, on the lines of the Myers' index of concentration, was evolved in this note, leading to an index of aggregate distortion; a relative range measure of deviation and a group efficiency index were suggested to enable better analysis and comparative study. It was interesting to note that the average deviation percent of age was nearly uniform in all age ranges, of the order of 0.5. A new technique was applied to determine the most efficient set of age grouping, as the group efficiency index varied for different age segments and socio-economic classes of the same population.

Section Two

THE PROBLEM

2.1. While grouping of data is often necessary for presentation and proper comprehension, in the field of age statistics this necessity may be utilised to evolve a set of grouping that reduces the total group errors to a minimum. In a country where ages are as a rule definitely known and reported, the question of the most efficient set of age grouping is not so important, as the group errors will be small for any set. But in a country where ages are generally not definitely known and the heaping up at certain digits at the cost of others is very marked, the selection of an efficient set of grouping is very important.

2.2. This feature of heaping up or concentration at certain popular digits was usually referred as 'integer bias' in the past and sought to be attributed to bias for certain 'preferred' end-digits like 0 and 5. In recent years, however, this is being treated more as an error of rounding off. A good deal of work on the subject of 'integer-bias' or 'round-off' errors in age reporting has already been done, specially in the national census publications of different countries. Age is an important factor not only in the understanding of the vital flows that condition population dynamics but also in sizing up most other population characteristics of socio-economic interest; the need for getting at the best estimate of the true group-age distribution is thus obvious.

2.3. *A priori* considerations suggest that the heaping up in age returns might be the combined effect of the following elements :

(1) digit preference (as such, without the effects of estimation error and age bias);

(2) estimation error (as such, without the effects of age bias);

(3) age bias.

An effort was, therefore, made to grasp the effect of these elements in isolation and to build up the most efficient set of grouping from the knowledge so obtained. The study of these forces in isolation was difficult and some amount of overlapping could not always be avoided.

2.4. From *a priori* considerations again, it would appear that the effect of digit-preference can extend over the unit cycle of end-digits 0, 1, 2, ..., 9 and thus only small displacement errors independent of the age range, should result from it. Estimation error could similarly be expected to produce displacements, small in the earlier age ranges but gradually increasing as age advances. The digit preference and the estimation error again, from their very nature, would be of cyclical nature over the array of end-digits. The age bias, arising as it does from extraneous influences was more likely to have a few focal points at the crucial ages (specific

for different countries in a given period of time), apart from some general tendency to understate or overstate at particular age regions, without any cyclical characteristics : the pull of age bias is apt to be lopsided and to have a long arm.

2.5. If ages are exactly known, stated and recorded the true distribution will of course be reproduced. When ages are exactly known but not correctly stated, age bias will obviously be the element responsible. If ages are known within a narrow margin, the digit preference may conceivably be an important element; but when ages are not known, or only known to lie within widely separated limits, the estimation error is likely to be the dominant element. It is natural that more than one element will be found superimposed on the dominant element in a practical situation, for example when the ages are only known to lie within widely separated limits, the limiting ages themselves will be liable to the influence of age bias.

2.6. What usually happens in a country like India is that the age is unknown and has to be estimated from looks or from comparison with known events or relative seniority ranking within the household or community. Such assessment of age has to be done by the field investigator or enumerator : in reality, an age band with its length depending on the type of evidence available, is consciously or unconsciously estimated by the investigator in the first instance and before the allocation to an individual age within the band. Behind each of the recorded individual ages (falling in the category not definitely known) is, therefore, an estimation age band.

2.7. In WBSD Study[2·1], among other things, information about the type of evidence available on age was collected, along with the age as stated by the informant, the rating of the statement and the age assessed by the investigator in the field. Information about the type of available evidence is set in Table (2.1).

2.8. It will be seen from Table (2.1) that in West Bengal, where the accuracy of age assessment might be the best for India[2·2], year of birth of only about 15 per cent of the total population was definitely known. The ages could be estimated or known approximately in about 40 per cent cases; and for the balance of about 45 per cent the age recorded was more or less guess-work estimates. Even in the definite class, documentary evidence of age was obtained in negligible proportion of cases, particularly in the city area. This position should be borne in mind while considering the age returns in the Indian situation.

2·1 This was an experiment on methodology conducted in connection with the NSS. The NSS 4th round sample villages, and the urban and city blocks in West Bengal were adopted for the Study. 744 sample households (hhs.) in 71 villages, 405 sample households in 26 urban blocks and 170 sample households in 14 city blocks (in Calcutta) were interviewed in the Study. Only 18 households had to be substituted, mostly occasioned by subsequent removal. Original NSS sampling fractions were adjusted in a manner as to give uniform multipliers for the three agglomeration sectors. 43 investigators were employed in the experiment and about 1850 investigation-inspection days used up during April-June in the Study.

2·2 As measured by the Index of Concentration evolved by the U.S. Bureau of the Census and adopted by the Indian Census; *Census of India 1951, Paper No. 3, 1954,* p.4.

TABLE (2.1): DISTRIBUTION OF INDIVIDUALS BY TYPE OF AVAILABLE EVIDENCE ABOUT AGE

(NSS WBSD Study 1954)

	age-assessed group	type of evidence				total
		hearsay, guess or eye-estimate	related with definite or approximate ages or events	definite statement of year of birth	birth certificate or other documentary	
	(1)	(2)	(3)	(4)	(5)	(6)
		city (170 households)				
1.	0— 6	26	49	21	—	96
	(%)	(27.1)	(51.0)	(21.9)		(100.0)
2.	7—16	40	40	16	—	96
	(%)	(41.7)	(41.7)	(16.6)		(100.0)
3.	17—above	262	79	54	3	398
	(%)	(65.8)	(19.8)	(13.6)	(0.8)	(100.0)
4.	all ages	328	168	91	3	590
	(%)	(55.6)	(28.5)	(15.4)	(0.5)	(100.0)
		other urban (405 households)				
1.	0— 6	81	138	105	5	329
	(%)	(24.6)	(42.0)	(31.9)	(1.5)	(100.0)
2.	7—16	137	199	60	8	404
	(%)	(33.9)	(49.3)	(14.8)	(2.0)	(100.0)
3.	17—above	581	423	60	20	1084
	(%)	(53.6)	(39.0)	(5.5)	(1.9)	(100.0)
4.	all ages	799	760	225	33	1817
	(%)	(44.0)	(41.8)	(12.4)	(1.8)	(100.0)
		rural (754 households)				
1.	0— 6	129	232	302	20	683
	(%)	(18.9)	(34.0)	(44.2)	(2.9)	(100.0)
2.	7—16	322	399	132	24	877
	(%)	(36.7)	(45.5)	(15.1)	(2.7)	(100.0)
3.	17—above	958	920	137	57	2072
	(%)	(46.2)	(44.4)	(6.6)	(2.8)	(100.0)
4.	all ages	1409	1551	571	101	3632
	(%)	(38.8)	(42.7)	(15.7)	(2.8)	(100.0)

2.9. Chart (1) gives the histogram representing the distribution of NSS 4th round all-India urban sample population in individual ages, to demonstrate the extent of age distortions involved. The seriousness of the situation will be apparent when it is realised that in the NSS material used in the histogram the population returned at the single individual age 60 is 53.4 per cent of the total population returned in the age group 56–60 and 64.5 per cent of the total returned in age group 60–64; the respective proportions for the rural sector are 61.1 per cent and 68.6 per cent. And the quality of age reporting in NSS medium, as would be anticipated, was somewhat superior to the general census standard. These facts on the type of evidence demonstrate how limited is the possibility of improving the quality of the Indian age returns, perhaps for at least a generation to come.

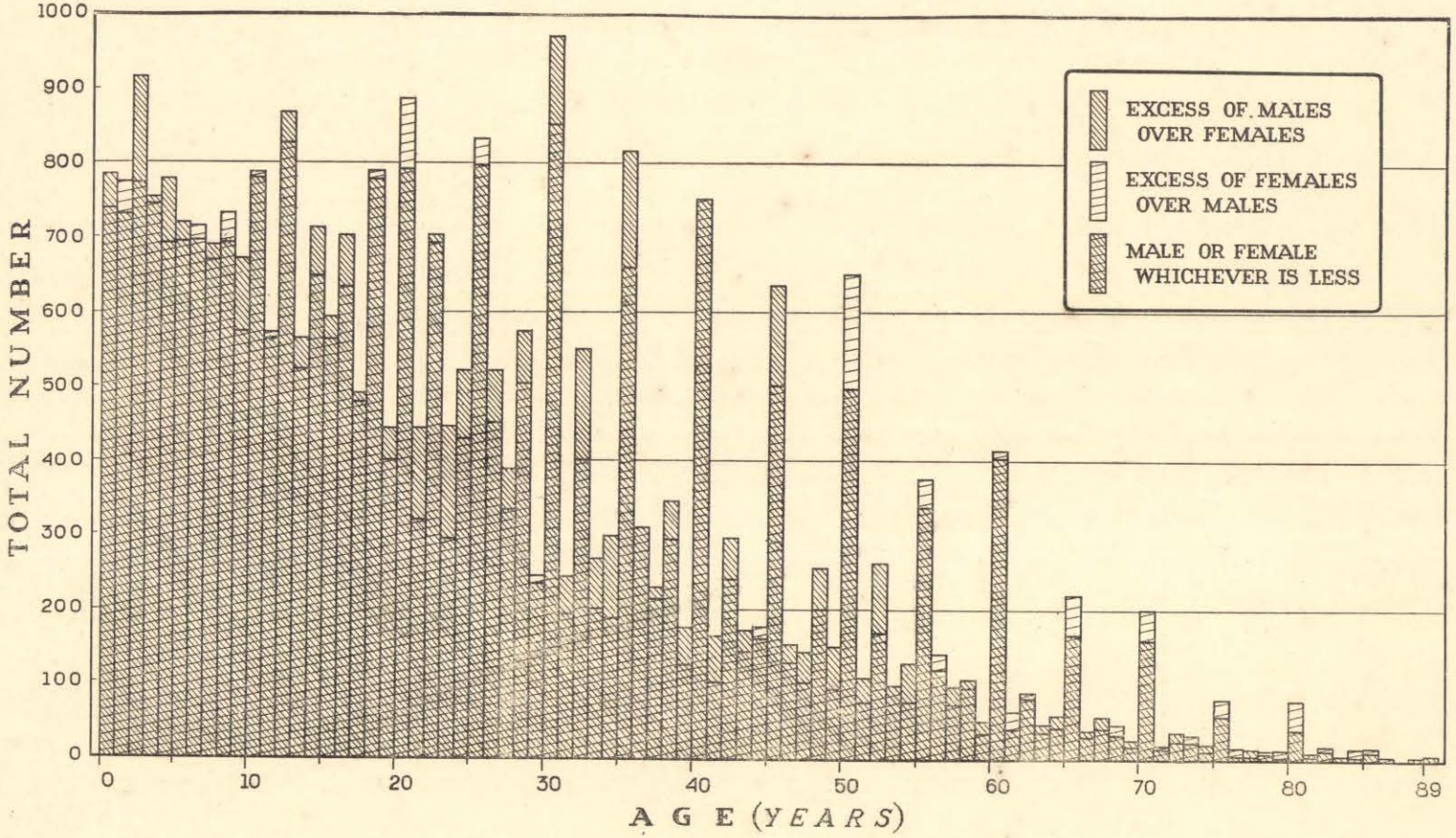


CHART (1) : Frequency distribution of total number returned at each individual age (NSS 4th round, urban).
The number of males or females returned are shown in the diagram on a vertical scale half of that for total numbers. The total population lies exactly half way between the top of the criss-cross portion and the top of the bar.

2.10. Discussions will be confined in this note to uniform quinary groups, the smallest groups usually adopted in representation of age statistics. Analysis for a most efficient set of decennial groups will involve similar considerations, though extension of the group interval to cover the full cycle of digits makes the problem somewhat easier here. Systems of unequal groups (alternate ternary—septanary for example) are unusual and inconvenient for presentation purposes. In what follows accordingly by the set of most efficient grouping will be meant the set of uniform quinary groups which gives the best fit to the true conditions over the whole range of life. The chronological age is only dealt with in this note.

2.11. As will be apparent from subsequent discussions a certain set of grouping may not be most efficient both in rural and urban conditions, and for different income slabs within the population. In the same manner, it is conceivable that one single set may not be equally efficient, say, for different economic activity segments like earners and dependants, or for different universe of events like births and deaths. As treatment of different sectors of the population in varying sets of grouping will disturb comparability, compromise is clearly called for.

Section Three

DIGIT PREFERENCE

3.1. The nature and extent of the digit preference was studied in isolation in the first instance. In the E & ESP Study 1954 conducted in the Indian Statistical Institute (ISI), a cluster of five 7-digited numbers without the 3 middle digits was given to the workers for supplying the suppressed middle digits by guess. The distribution of the central digit so supplied by 220 workers, expected to be free from any extraneous bias apart from the digit preference as such[3·1], is given in Table (3.1).

TABLE (3.1): FREQUENCY DISTRIBUTION OF THE CENTRAL MISSING DIGIT SUPPLIED BY GUESS

(ISI, E and ESP Study 1954)

	digit										total
	0	1	2	3	4	5	6	7	8	9	
frequency	86	59	133	117	115	134	127	140	94	95	1100
(%)	(7.8)	(5.4)	(12.1)	(10.6)	(10.4)	(12.2)	(11.6)	(12.7)	(8.6)	(8.6)	(100.0)
with expected frequency 110 in each cell, $\chi^2 = 54.8$, significant at 1%.											

3.2. It was apparent that true digit preference comprised a tendency to keep to the middle of the digit array 0, 1, ..., 9, within the range 2–7. The shortfall of the actual frequency from the expected (110) was quite marked at the end-digits, 0, 1, 8, 9[3·1]. This pattern of digit preference is altogether different from the 'integer bias' with strong pulls for 0 and 5 and smaller pulls for 2 and 8 that emerges from the analysis of age returns. As will be seen later, the heaping up at 0, 5, 2 and 8 arises mainly from estimation error, which is the most powerful element behind the distortion in age reporting.

3.3. Other interesting facets of digit preference were disclosed by the E & ESP Study[3·1]. One was the disinclination to repeat digits in one sequence and the other the preference for a run of consecutively rising digits; we shall call this the second order of digit preference. Table (3.2) gives the distribution of two consecutive filled-in digits in the E & ESP study (the missing third and fourth places of 7-digited numbers).

3·1 It is relevant to mention that the effect of extra-sensory perception (ESP) was found negligible. Incidentally, an understanding of digit preference may be usefully employed to check and control (numerical) copying and computational mistakes.

TABLE (3.2): FREQUENCY DISTRIBUTION OF DIGITS SUPPLIED BY GUESS IN THE FIRST TWO CONSECUTIVE MISSING DIGIT PLACES
(ISI, E and ESP Study 1954)

first missing digit	second missing digit										
	0	1	2	3	4	5	6	7	8	9	total
0	*9*	**7**	9	4	6	5	1	2	3	3	49
1	14	*9*	**29**	13	10	10	7	7	6	3	108
2	12	8	*6*	**22**	16	13	18	12	9	7	123
3	6	4	14	*6*	**33**	20	15	18	11	14	141
4	11	4	19	24	*5*	**38**	32	13	7	8	161
5	9	5	12	14	13	*9*	**24**	17	14	6	123
6	5	5	16	14	7	10	*7*	**40**	11	7	122
7	5	5	7	5	7	11	13	*7*	**21**	13	94
8	5	3	11	6	9	10	4	16	*3*	**21**	88
9	10	9	10	9	9	8	6	8	9	*13*	91
total	86	59	133	117	115	134	127	140	94	95	1100

3.4. Table (3.3) shows separately the frequencies of selected digit pairs of Table (3.2). The expected frequency in each cell of the table on basis of random selection of digits is 11. Some of the lowest frequencies occurred with the repeated digit pairs 00, 11, 22, ..., 99, the total frequency of this set of ten repeated digit numbers being only 74 against expected 110. On the other hand, most of the highest ferequencies occurred with the set of eight consecutively rising digit pairs 12, 23, ..., 89, the total frequency of the set being 228 against expected 88.

TABLE (3.3): FREQUENCY DISTRIBUTION OF SELECTED PAIRED CONSECUTIVE DIGITS SUPPLIED BY GUESS
(ISI, E and ESP Study 1954)

consecutive rising run of digits		repeated digits	
selected paired digits	frequency	selected paired digits	frequency
(1)	(2)	(1)	(2)
12	29	00	9
23	22	11	9
34	33	22	6
		33	6
45	38	44	5
56	24	55	9
67	40	66	7
		77	7
78	21	88	3
89	21	99	13
total	228	total	74
average	28.5	average	7.4
with expected frequency 11 in each cell $\chi^2 = 263.4$ with degrees of freedom 8, significant at 1%.		with expected frequency 11 in each cell $\chi^2 = 18.2$ with degrees of freedom 10, significant at 5%.	

3.5. Similar preferences for run of consecutively rising digits and dislike for the repeated digits were found in the analysis of other filled-in places of the numbers. Part of the short-fall in frequencies of the repeated digit pairs clearly resulted from the attraction for the consecutively rising digit pairs, contiguous to them. The preference is also disclosed in the frequency distribution of all the three filled-in missing digits given in Table (3.4). Arranged in the table in the order in which they occurred in the filled-in E & ESP schedules, the progressive diagonal shift of the maximum frequency range down the table is apparent.

TABLE (3.4): FREQUENCY DISTRIBUTION OF ALL THE THREE CONSECUTIVE MISSING DIGITS SUPPLIED BY GUESS

(ISI, E and ESP Study 1954)

digit	third place	fourth place	fifth place
(1)	(2)	(3)	(4)
0	49	86	85
1	108	59	90
2	123	133	88
3	141	117	139
4	161	115	95
5	123	134	143
6	122	127	118
7	94	140	102
8	88	94	116
9	91	95	124

3.6. These inherent likes and dislikes in the run of numbers will naturally enter the reporting by the informant and the assessment and recording by the investigator. The reported individual age distribution of Bengal (males) of Census 1911 and of West Bengal and U.P. (males) of Census 1951 were examined to see how far this second order of digit preference persisted in the census age reporting. The disturbance in age returns from estimation error was much stronger, but the digit preference of the second order being non-cyclic was not masked altogether by this stronger cyclic disturbance. The numbers returned at particular ages under examination could not be compared with the graduated frequency for the purpose of the examination and a method had to be devised to estimate the expected frequency on elimination of the second order of digit preference.

3.7. Chart (2) shows for Bengal (males) 1911 Census population, in the form of smooth curves, the distribution of the numbers returned in individual ages in decennial age-segments.

3.8. It was argued that the form of the frequency curve in the region of an individual age in question, subject to the primary digit preference and estimation error (both of which were of cyclical nature within the digit array) but not subject to the second order of digit preference, will be intermediate to the form of the curve in similar end-digit regions either side, ten years of age up and below. In other

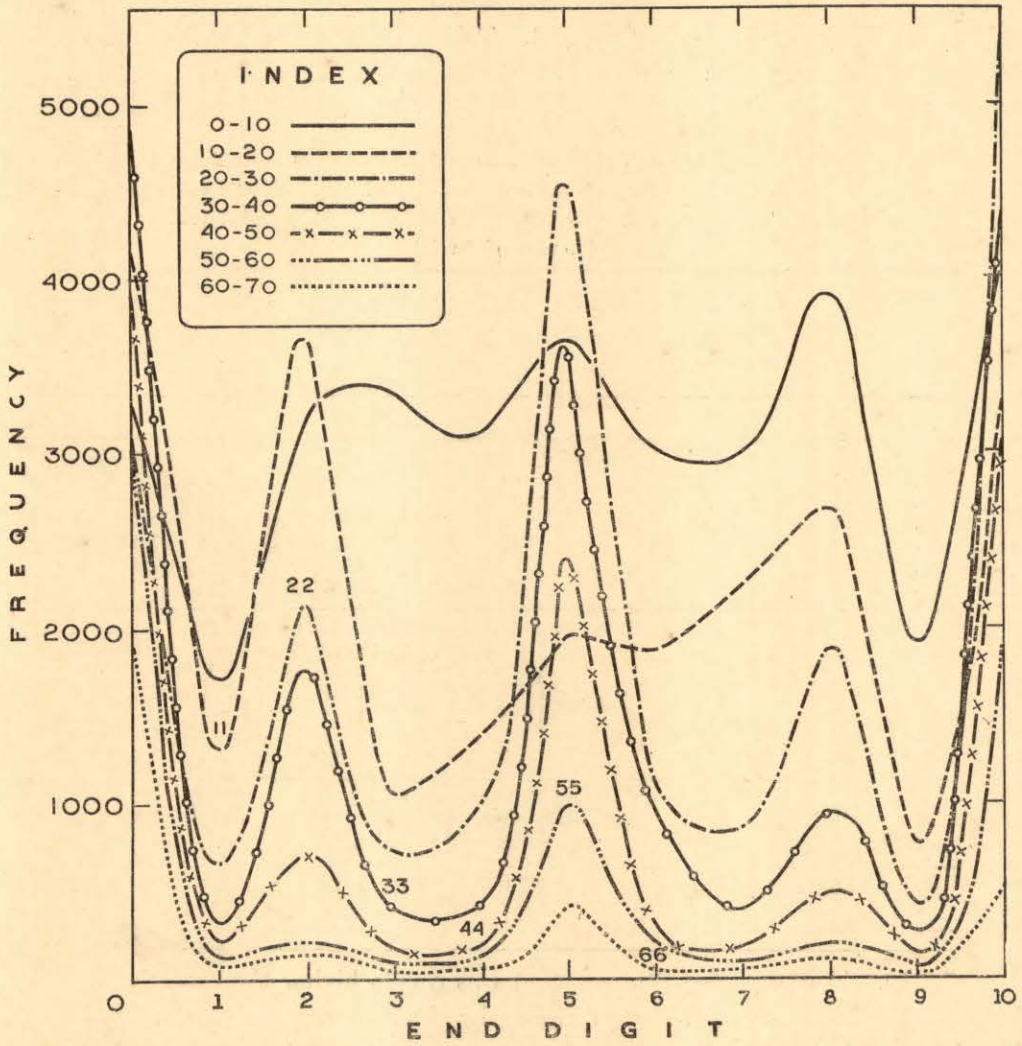


CHART (2) : Frequency curves of numbers returned at each end-digit in decennial age groups (Census 1911, Bengal males).

words, denoting the actual number returned at individual age x, by the symbol n_x the form of the individual age frequency curve in the $n_{10v+u-1} : n_{10v+u} : n_{10v+u+1}$ region would be intermediate between those rendered by the $n_{10(v-1)+u-1} : n_{10(v-1)+u} : n_{10(v-1)+u+1}$ and $n_{10(v+1)+u-1} : n_{10(v+1)+u} : n_{10(v+1)+u+1}$ regions, if the second order of digit preference were not there. In effect the assumption allows only for the cyclic distortions and thus eliminates the non-cyclic influences. The distortion from age bias is therefore also not allowed for ; but for simplicity the age bias which is localised can be left out of account for the time being.

3.9. The simplest estimations were made in pursuance of the above assumption and constants determined from two linear simultaneous equations on either side were applied to the repeated digit age in question to estimate the appropriate expected frequency. Thus, to estimate $E(n_r)$ the expected frequency at the repeated digit age $r = 10v + v$ the formula $E(n_r) = A_v n_{r-1} + B_v(n_r + n_{r+1})$ was tried, the constants A_v, B_v being determined from the two equations $n_{r-10} = A_v n_{r-10-1} + B_v(n_{r-10} + n_{r-10+1})$ and $n_{r+10} = A_v n_{r+10-1} + B_v(n_{r+10} + n_{r+10+1})$. Table (2.4) gives the numbers actually returned and the numbers expected at the repeated digit ages, for Census 1911 Bengal (males), and Census 1951 West Bengal (males) and Uttar Pradesh (males).

TABLE (3.5): POPULATION RETURNED AT REPEATED DIGIT INDIVIDUAL AGES IN CENSUS AND EXPECTED POPULATION ON ELIMINATION OF SECOND ORDER OF DIGIT PREFERENCE

(Census of India)

	age x	number returned in census (000) n_x	expected frequency (000) $E(n_x)$	percentage deviation $\frac{E(n_x)-n_x}{n_x}\times 100$
	(1)	(2)	(3)	(4)
		Bengal (males) : Census 1911		
1.	11	1310	1606	22.6
2.	22	2156	2304	6.9
3.	33	374	363	−2.9
4.	44	202	218	7.9
5.	55	1017	1096	7.8
6.	66	39	41	5.1
		West Bengal (males) : Census 1951		
1.	11	2395	2799	16.9
2.	22	2547	2474	−2.9
3.	33	1397	1491	6.7
4.	44	1250	1292	3.4
5.	55	1068	1119	4.8
6.	66	212	232	9.1
		U. P. (males) : Census 1951		
1.	11	6408	8731	36.3
2.	22	6356	6112	−3.8
3.	33	2073	3003	44.9
4.	44	1871	1918	2.5
5.	55	4991	5332	6.8
6.	66	374	406	8.6

3.10. The expected frequency was as a rule higher than numbers returned; the expected frequency was always higher for the repeated digit pairs age 44 above, though small derivations in reverse direction appeared in the younger repeated digit pair ages. As stated earlier, the age bias could not be allowed for in the method of estimation used and the element of age bias perhaps disturbed the expected frequency of the earlier repeated digit pair ages rendered by the method. Use of a broader based formula might have given better balanced estimates. But the evidence in support of the hypothesis that the second order digit preference persists in age reporting was clear enough from the analysis done above.

3.11. It seems probable that the inflation noticed at age 60 in the age returns of many countries may be, at least partly, due to the dislike of the repeated digit number 55.

Section Four

ESTIMATION ERROR

4.1. Some results obtained in the E & ESP Study relevant to the estimation error are discussed first. The E & ESP Study schedule also contained a cluster of five lines, of which the lengths were required to be eye-estimated and recorded to the second place of decimal in terms of an unconventional unit of length 'L' shown on the body of the schedule[4·1]. The conditions were such that the second decimal figure could be nothing better than pure guess. The distribution of the second decimal figure supplied by 222 workers is given in Table (4.1).

TABLE (4.1): DISTRIBUTION OF (1) THE SECOND PLACE AFTER DECIMAL OF THE EYE-ESTIMATED LENGTH OF LINES AND (2) THE END-DIGIT OF AGE OF ALL-INDIA RURAL SAMPLE POPULATION AGED 40-ABOVE

(ISI, E and ESP Study 1954 and NSS 4th round[4·2] 1952)

item	digit										total
	0	1	2	3	4	5	6	7	8	9	
1. second decimal place of estimate	414	25	72	45	38	344	58	41	48	25	1110
(%)	(37.3)	(2.3)	(6.5)	(4.0)	(3.4)	(31.0)	(5.2)	(3.7)	(4.3)	(2.3)	(100.0)
2. end digit, age 40-above	2326	276	552	279	297	1265	301	234	352	165	6047
(concentration)[4·3]	(31.3)	(4.3)	(8.8)	(4.7)	(5.2)	(22.7)	(6.2)	(5.1)	(7.8)	(3.9)	(100.0)

4.2. The concentration at 'preferred' digits is now of the familiar pattern found in age returns, though more accentuated here for the digits '5' and '0'. The striking similarity between the frequency distribution of the digit in the second decimal place in the estimated lengths of lines in the E & ESP Study and the end digit of age of the all-India rural sample population aged 40 and above, suggested that most of the error in age returns is that of estimation when the unit digit of age was just a matter of guess.

4.3. On further analysis, a tendency to over-estimate was also disclosed by the E & ESP Study which was suggestive. The actual aggregate length of the five lines correct to the first place of decimals was 6.3L; but the mean of the estimates recorded by the workers was 6.7L. The distribution of the recorded estimates is given in Table (4.2). The over-estimation was highly significant; as against only 27% who came on the side of under to correct estimation, 73% over-estimated the aggregate length. The major peak of the distribution of estimates was at 6.6L.

4·1 Appendix 0.

4·2 All the NSS 4th round rural tables of this note cover the six 1/16th part samples 3, 4, 7, 8, 15 & 16 split at the village level for operational convenience.

4·3 The measure of concentration is a percentage distribution of the end digits defined in para 6.5.

TABLE (4.2): DISTRIBUTION OF EYE-ESTIMATE OF THE AGGREGATE LENGTHS OF A CLUSTER OF 5 LINES (ACTUAL AGGREGATE 6.33L) ROUNDED TO THE FIRST DECIMAL PLACE

(ISI, E and ESP Study 1954)

	length (L)	frequency		length (L)	frequency		length (L)	frequency
	(1)	(2)		(1)	(2)		(1)	(2)
1.	5.1	1	11.	6.4	16	22.	7.5	2
2.	5.4	1	12.	6.5	20	23.	7.6	4
3.	5.7	1	13.	6.6	29	24.	7.7	1
4.	5.8	1	14.	6.7	11	25.	7.8	2
5.	6.0	11	15.	6.8	15	26.	7.9	3
6.	6.1	6	16.	6.9	14	27.	8.0	2
7.	6.2	15	17.	7.0	16	28.	8.1	1
8.	6.3	24	18.	7.1	7	29.	8.2	1
			19.	7.2	7	30.	8.3	1
9.	sub-total: 6.3 below	60	20.	7.3	3	31.	8.4	1
	(%)	(27.0)	21.	7.4	6			
10.	sub-total: 6.4 above	162				32.	total	222
	(%)	(73.0)					(%)	(100)

mean = 6.69 $\quad \sigma^2 = .2627 \quad s_m = .0434 \quad t = 8.3$

4.4. Such tendency of over-estimation (or under-estimation) has been found by other operators in different experiments[4·4].

4.5. It was therefore decided to investigate how far the bias to over-estimate or under-estimate might have entered the assessment of age. Actual age data of the WBSD Study were analysed to investigate this. Table (4.3) gives the distribution of the individuals covered by the Study in age-assessed minus age-stated groups, separately for the city, other urban and rural sectors[4·5].

TABLE (4.3): DISTRIBUTION OF INDIVIDUALS IN AGE-ASSESSED MINUS AGE-STATED CLASSES, UNDER EDUCATION STANDARD BREAKDOWNS

(NSS, WBSD Study 1954)

age-assessed minus age-stated	city (170 households)				other urban (405 households)				rural (754 households)			
	total	illi-terate	liter-ate	matri-culate	total	illi-terate	liter-ate	matri-culate	total	illi-terate	liter-ate	matri-culate
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1. age-assessed = age-stated	495	194	236	65	1552	749	710	93	2733	2012	701	20
(%)	(83.9)	(86.2)	(81.9)	(84.4)	(86.2)	(81.6)	(90.5)	(94.9)	(77.6)	(76.3)	(81.1)	(95.2)
2. age-assessed < age-stated	20	8	10	2	70	59	11	—	274	218	56	—
(%)	(3.4)	(3.6)	(3.5)	(2.6)	(3.9)	(6.4)	(1.4)	(—)	(7.8)	(8.3)	(6.5)	(—)
3. age-assessed > age-stated	75	23	42	10	178	110	63	5	515	407	107	1
(%)	(12.7)	(10.2)	(14.6)	(13.0)	(9.9)	(12.0)	(8.1)	(5.1)	(14.6)	(15.4)	(12.4)	(4.8)
4. total	590	225	288	77	1800	918	784	98	3522	2637	864	21
(%)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)

[4·4] Examples of such bias of over-estimation in selection of 'representative' units have been cited by Frank Yates in "*Sampling Methods for Censuses and Surveys*" (1953) at pp. 12–13.

[4·5] Age was not stated in less than 1% of the cases only (most of it in the rural sector) and the not-stated cases have been left out of the tables.

4.6. The age-assessed minus age-stated was positive two to four times more often than it was negative : that is, a higher age was assessed two to four times more frequently. With regard to this feature the male female differential was not significant, but the relative proportion of higher age-assessed tended to go up among non-Hindus and among Hindi-speaking population in the West Bengal field[4·6]. The proportion, was, if anything, rather higher in city area and among the educated, as Table (4.3) shows.

4.7. The investigator's assessment of age is usually the only thing available and recorded in census and surveys in India but in the WBSD Study the investigators were definitely instructed not to render any such assistance in age statement and clear evidence is thus furnished that the investigator assesses a higher age in the sum than the informant states. The question is whether the investigator was trying to correct in his assessment a real under-statement of age by the informant, or if in the aggregate there was no such under-statement in the process of assessment, the age was over-estimated by the investigator. An attempt was made to answer this question from further examination of the WBSD Study material. Table (4.4) gives the distribution of individuals in age-assessed minus age-stated classes for different categories of rating of statement.

TABLE (4.4) : DISTRIBUTION OF INDIVIDUALS IN AGE-ASSESSED MINUS AGE-STATED CLASSES UNDER RATING OF STATEMENT CATEGORIES
(NSS, WBSD Study 1954)

age-assessed minus age-stated	rating of age statement								
	city (170 households)			other urban (405 households)			rural (754 households)		
	guess	approximate	difinite	guess	approximate	definite	guess	approximate	definite
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. age-assessed =age-stated	180	153	162	350	513	689	733	992	1008
(%)	(79.3)	(81.8)	(92.0)	(74.2)	(82.7)	(97.3)	(63.3)	(75.4)	(96.3)
2. age-assessed < age-stated	11	5	4	38	28	4	142	120	12
(%)	(4.8)	(2.7)	(2.3)	(8.0)	(4.5)	(0.6)	(12.2)	(9.1)	(1.1)
3. age-assessed > age-stated	36	29	10	84	79	15	284	204	27
(%)	(15.9)	(15.5)	(5.7)	(17.8)	(12.8)	(2.1)	(24.5)	(15.5)	(2.6)
4. total	227	187	176	472	620	708	1159	1316	1047
(%)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

4.8. For all categories of rating, the proportion of ages assessed higher to ages assessed lower is fairly stable, about 2 for all sectors combined, rather a little higher for the category of rating definite. A progressive fall in the proportion was to be expected in passing from guess to definite category of rating of age statement if the investigator, in his assessment, was correcting under-statements for which the guess and approximate categories offered a much bigger scope. The actual pattern brought out suggests over-estimation in age-assessed.

4.9. The assumption is, however, implicit here that the rating of statement has been proper. As to the reliability of the rating, Table (4.5) gives the frequency

[4·6] Appendix 1.

of end-digit '0' in the age-stated range 23–62; a systematic fall in the concentration at end-dight '0' with upgrading in rating category is observed. It is permissible to take this as a very rough test of the validity of rating done.

TABLE (4.5): CONCENTRATION AT END-DIGIT '0' IN AGE STATEMENTS UNDER DIFFERENT RATING OF STATEMENT CATEGORIES

(NSS, WBSD Study 1954)

rating of statement	city (170 households)			other urban (405 households)			rural (754 households)		
	population aged 23-62		concentration	population aged 23-62		concentration	population aged 23-62		concentration
	returned at end-digit '0' ages	total	col(2) / col(3) ×100	returned at end-digit '0' ages	total	col(5) / col(6) ×100	retrurned at end-digit '0' ages	total	col(8) / col(9) ×100
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. guess	37	161	23.0	71	226	31.4	160	596	26.9
2. approximate	14	84	16.6	70	341	20.5	143	665	21.5
3. definite	5	51	9.8	36	197	18.3	39	219	17.8
4. total	56	296	18.9	177	764	23.2	342	1480	23.1

4.10. Table (4.6) gives distribution of the gap between age-assessed and age-stated in broad age-assessed groups.

TABLE (4.6): DISTRIBUTION OF INDIVIDUALS IN DIFFERENT AGE RANGES UNDER AGE-ASSESSED MINUS AGE-STATED GROUPS

(NSS, WBSD Study 1954)

age-assessed minus age-stated	age-assessed (years)								
	city (170 households)			other urban (405 households)			rural (754 households)		
	0–16	17–61	62–above	0–16	17–61	62–above	0–16	17–61	62–above
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. –11 below	–	–	1	–	2	–	–	9	2
2. –10 to –6	–	1	–	–	3	–	1	28	2
3. – 5 to –3	–	2	2	5	14	3	13	58	8
4. – 2 to –1	5	8	1	20	23	–	73	76	4
5. –1 & below	5	11	4	25	42	3	87	171	16
(%)	(2.6)	(2.9)	(16.0)	(3.4)	(4.2)	(4.8)	(5.7)	(9.1)	(14.7)
6. 0	182	296	17	684	815	53	1363	1306	64
(%)	(94.8)	(79.4)	(68.0)	(93.3)	(81.1)	(85.5)	(88.6)	(69.7)	(58.7)
7. 1 to 2	5	42	2	20	82	–	75	159	6
8. 3 to 5	–	21	2	4	54	6	13	183	13
9. 6 to 10	–	3	–	–	8	–	–	49	8
10. 11 above	–	–	–	–	4	–	–	7	2
11. 1 & above	5	66	4	24	148	6	88	398	29
(%)	(2.6)	(17.7)	(16.0)	(3.3)	(14.7)	(9.7)	(5.7)	(21.2)	(26.6)
12. total	192	373	25	733	1005	62	1538	1875	109
(%)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

4.11. While a tendency to overstate age in the old age range is well known and is likely to have brought down the proportion of the total over-assessment to the total under-assessment in the old age range 62-above, the even break of over-assessment and under-assessment in the young age range 0–16 is not so easily explained : the margin available for under-statement is however limited in the young age range by the age attained. The spread up of the gap between age-assessed and age-stated is interesting; less than half of the deviations exceed 2 years of age and most of it fall within the limits of ± 5 years.

4.12. In the WBSD Study, as already indicated, information was collected about the type of evidence available to the investigator in assessment of ages, on his best efforts. Table (4.7) is an alternative presentation of the information obtained in this respect; both the rating of assessment and type of evidence on which the assessment naturally rested were combined here to give composite categories for assessment evidence types. Table (4.7) shows that definite evidence of age was available in 18–30% of cases only; and, as was seen from Table (2.1), a definite statement of age by the informant was behind most of it. Considering that 16–19% of the individuals covered were in the age range 0–6, the grave weakness in the field of the age assessment is quite apparent. The proportions with definite assessment-evidence type were higher than the respective proportions with definite evidence of age available given in Table (2.1), and the gap increased progressively in passing from the city to the rural sector. The age-assessed series could not be taken as a better approximation to the true ages, in the circumstances.

TABLE (4.7) : DISTRIBUTION OF INDIVIDUALS IN ASSESSMENT-EVIDENCE TYPE CATEGORIES UNDER SEX BREAKDOWNS

(NSS, WBSD Study 1954)

assessment-evidence type	city (170 households)			other urban (405 households)			rural (754 households)		
	total	male	female	total	male	female	total	male	female
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. guess	276	154	122	372	204	168	883	434	449
(%)	(46.8)	(45.2)	(49.0)	(20.5)	(21.0)	(19.8)	(24.3)	(23.2)	(25.5)
2. approximate	191	108	83	1025	521	504	1743	876	867
(%)	(32.4)	(31.7)	(33.3)	(56.4)	(53.8)	(59.4)	(48.0)	(46.7)	(49.3)
3. definite	123	79	44	420	244	176	1006	564	442
(%)	(20.8)	(23.1)	(17.7)	(23.1)	(25.2)	(20.8)	(27.7)	(30.1)	(25.2)
4. total	590	341	249	1817	969	848	3632	1874	1758
(%)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

4.13. The values of concentration at end digit '0' under different ratings of assessment, similar to those shown in Table (4.5) but for the age-assessed series now, are given in Table (4.8). Comparative study of the concentration makes it clear that the quality of the age-assessed series is no better than the age-stated series.

TABLE (4.8): CONCENTRATION AT END-DIGIT '0' IN AGE-ASSESSED SERIES UNDER DIFFERENT RATING OF ASSESSMENT CLASSES

(NSS, WBSD Study 1954)

rating of assessment	city (170 households)			other urban (405 households)			rural (754 households)		
	population aged 23-62 returned at end digit '0' ages	population aged 23-62 total	concentration col(2)/col(3)×100	population aged 23-62 returned at end digit '0' ages	population aged 23-62 total	concentration col(5)/col(6)×100	population aged 23-62 returned at end digit '0' ages	population aged 23-62 total	concentration col(8)/col(9)×100
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. guess	39	150	26.0	49	207	23.7	128	442	29.0
2. approximate	14	101	13.9	97	437	22.2	204	889	23.0
3. definite	3	46	6.5	26	140	18.6	44	273	16.1
4. total	56	297	18.9	172	784	21.9	376	1604	23.5

4.14. The probable reason why '0' and '5' happen to be the most favoured digits, in that order, can be examined here. Rounding off at '0' naturally gets first preference as one digit, the unit place, is cut out by such approximation; and at that level of approximation, the mid-way digit '5' gets the next natural preference when the estimate is far off from the rounded up ages at the tenth place on either side. The digit preference as such may also be partially responsible for the popularity for the middle of the array digit '5'. After '0' and '5' there is usually a slight preference for even digit over odd: this also has a simple explanation. If an estimate is above '0' but not far away removed from '0', it will be transferred and recorded under '0' and not under '1'; only when it is far away removed from '0' it will be recorded differently, and will then rather be jumped to '2'. The preference for digit '8' has exactly a similar explanation.

4.15. "Census of India 1951—Age Tables"[4.7] gives a diagonal Table for the Madras (male) population showing estimated percentage 'under-statements' and 'over-statements' at each age from 6 to 67. This Table envisages notional transfers only to adjoining ages, and the graduated individual age frequencies taken as true are derived from the set of grouping centred round the most preferred digits '0' and '5'. The set of grouping does not take into account the tendency of over-estimation. In the age range 27–67 of this table, the average over-statement of age is 29 per cent

4.7 *Census of India*, 1951, *Paper No. 3, 1954*, pp. 16-18.

as against the average under-statement of about 20 per cent. In the age range 6–67, the proportions of similar average over-statement and under-statement are 22 per cent and 19 per cent respectively. Supplementary evidence of the tendency of over-estimation in age return is thus disclosed by the Census reporting itself.

4.16. The analysis done above shows that the estimation error really falls into two parts, one arising from rounding off approximation and other from over-estimation. The element of over-estimation missed attention in the past and the estimation error was equated to the error of rounding off.

4.17. In NSS experimental West Bengal Household Comparative (WBHC) Study 1955 the same sample households as of NSS 4th round were revisited after a lapse of about 3 years to measure changes in living conditions during the intervening period : this opportunity was utilised to investigate further the over-estimation bias in age reporting and the ages of the household members were independently ascertained again. Comparisons between ages reported in NSS 4th round and WBHC Study showed some interesting features. Table (4.9) gives the distribution of the deviation between the age reported in the WBHC Study and the age expected on the basis of the three-year old NSS 4th round age return.

TABLE (4.9) : FREQUENCY DISTRIBUTION OF THE NUMBER IN DIFFERENT AGE GROUPS BY ADJUSTED DIFFERENCE IN AGES

(NSS 4th round 1952 and WBHC Study 1955 : 650 rural households)

	age difference : WBHCS − (4th round + 3)	age (years) WBHCS					
		3—9	10—19	20—29	30—39	40—above	all ages
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1.	0	184	158	128	84	116	670
	(%)	(45.8)	(36.9)	(31.8)	(27.5)	(23.0)	(32.8)
2.	1 & above	91	126	134	125	244	720
	(%)	(22.6)	(29.4)	(33.2)	(41.0)	(48.4)	(35.3)
3.	mean difference	1.28	1.81	2.32	2.46	3.17	2.41
4.	−1 & below	127	144	141	96	144	652
	(%)	(31.6)	(33.7)	(35.0)	(31.5)	(28.6)	(31.9)
5.	mean difference	−1.67	−1.61	−1.95	−2.54	−2.79	−2.09
6.	total	402	428	403	305	504	2,042
	(%)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)
7.	mean difference	−0.24	−0.01	0.09	0.21	0.74	0.18
	with mean difference 0.18,	$t = 3.23$,	significant at 1%.				

4.18. The ages reported in WBHC Study were in sum somewhat higher than the ages expected on basis of the three-year old NSS 4th round returns, with an average over-statement of 0.18 years. But this was for the persons common in the two surveys : those born since NSS 4th round survey and those dead since, were naturally excluded from the comparison, apart from the migrants. The average report-

ed ages of the two surveys did not differ significantly. It was interesting to observe that while there was a big comparative under-statement in the youngest present age range 3–9, the difference narrowed in the age range 10–19, then changed to slight overstatement in the next higher age range 20–29 and to increasing over-statements in the later age ranges 30–39 and 40-above. The over-statement observed for the total range was thus the contributory effect of high over-statements at the advanced ages.

4.19. The investigating staff employed in both the surveys were by and large similar in instruction, training and experience; there were therefore no reasons to anticipate any material investigator differences. The reporting population was also more or less the same : they were not subject to repeated surveys in the intervening period nor otherwise conditioned to change. The WBHC Study results thus suggest that there is some under-estimation of advance of ages in the young age ranges below 20 and distinct over-estimation of the advance at the later adult ages, with a moderate over-estimation of ages in the balance. Thus in the aggregate the age of the population is over-estimated : the aggregate over-estimate may remain stable if the old, whose ages were over-estimated most in the previous survey, die in such proportion as to offset the subsequent increased over-estimation of those who live to grow older.

Section Five

AGE BIAS

5.1. The possible location and nature of the age bias in different population formations can often be known from their cultural traits. Tendencies to exaggerate ages in the threshold of adulthood and after retirement, conscious understatement of age in the young-adult range by the females in some culture and conscious over-statement of age to attain legal majority, to escape military service and to qualify for old-age pensions or just to impress as outstandingly old, are some of the common sources of the bias experienced. The mores and the laws of the land work behind the bias, and changes in them deflect the pattern of the bias : the pattern is usually quite stable over time in each country until the relevant laws change. The age bias is of specific location and is thus distinguished from the general estimation bias, from which it also differs in nature.

5.2. Longitudinal comparisons across consecutive census intervals, allowing for migration if significant, and for the digit preference and the estimation error, might disclose the pattern of age bias : except for the digit preference of the second order, these preferences and errors of cyclical nature get automatically allowed for to a large extent in comparisons over a series of decennial censuses. If reliable birth-death registration and migration statistics be available, the total distortions could be determined by reconciliation of successive census results with the intervening movements and the extent of age bias broadly assessed. Such techniques are however not helpful when the relevant data are grossly defective, as in India. Sample verification of ages in the field with a superior set of investigators, who travel down to the birth certificate or other best available evidence of age, is another alternative method employed to locate age bias or rather the total distortions in age recording. It may be reiterated that the cardinal point of interest in the problem of age grouping is the total distortion, and the elements leading to it are studied to get a better understanding of the position.

5.3. It should have been possible to spot the age bias at analysis stage by examining the run of the ratios that the numbers returned at each end digit of age constitute of the total returned in the successive decennial age ranges say. But such examination is also not very helpful for the Indian situation where the big distortions from estimation error mask most other features. Table (5.1) shows for the NSS medium the ratios

$$\frac{n_{10v+u}}{\sum_{u=0}^{9} n_{10v+u}} \times 100$$

up to age 79, for each end digit u in the whole sequence of age ranges

$$\sum_{u=0}^{9} n_{10v+u}\,, \qquad v = 0, 1, \ldots, 7.$$

TABLE (5.1): RATIO OF NUMBERS RETURNED AT EACH END-DIGIT TO TOTAL NUMBERS IN THE SUCCESSIVE DECENNIAL AGE RANGES

(NSS 4th round 1952, All-India rural sample: 28,918 persons)

end digit	decennial age range							
	0—9	10—19	20—29	30—39	40—49	50—59	60—69	70—79
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	10.6	15.2	18.7	26.6	31.4	38.6	49.2	51.0
1	10.1	8.1	6.1	4.6	4.5	5.0	4.6	3.8
2	10.9	14.4	13.5	11.8	9.5	9.2	8.9	7.7
3	10.8	8.7	6.9	5.8	5.0	4.5	3.8	4.4
4	9.7	9.4	9.1	5.8	5.6	4.6	4.7	3.0
5	10.9	10.4	18.4	20.4	23.1	19.6	18.7	19.7
6	10.3	11.3	8.4	8.3	6.2	5.0	2.7	3.3
7	8.9	5.8	5.4	4.5	4.9	3.8	2.1	2.7
8	10.2	11.7	9.8	8.4	6.8	6.5	3.6	3.0
9	7.6	5.0	3.7	3.8	3.0	3.2	1.7	1.4
total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

5.4. The ratios for any one end digit of age could have been expected ordinarily to form a smooth progression over the successive decennial age ranges, on which the impact of the age bias (ignoring the comparatively small influence of the second order of the digit preferences) will have produced marked local disturbances. But in any case, the history of the particular population growth and the local mores and laws to be turned to for confirmation : past fluctuations in birth-death experience may also produce isolated tracts of accumulation, particularly in small populations. Sharp rises in the ratios at ages 25 and 60 are observable in Table (5.1). The pull for age 60 was serious enough in the NSS medium to take the quinary age group 60–64 total beyond the 55–59 total. The examination of the run of the ratios as a method of spotting possible age bias is not satisfactory when the concentration at particular end digits is so high and the relative concentrations between the end digits change so violently as in Table (5.1).

5.5. The moot question in the present case was whether these sharp rises in the run of ratios came rather from mounting concentration at these preferred end digits. The age bias operated in a manner so as to accelerate and decelerate the flow of numbers returned in the immediate neighbourhoods of certain crucial ages : yet another approach of spotting the location of age bias suggested itself from this. Examination of the first differences of the ratios over a few consecutive end digits in neighbouring decennial age ranges should show up the acceleration and deceleration effects. Table (5.2) gives the first differences of the ratios in Table (5.1).

TABLE (5.2) : FIRST DIFFERENCES OF THE RATIOS OF NUMBERS RETURNED AT EACH END-DIGIT AS SHOWN IN TABLE (5.1)

(NSS 4th round 1952, all-India rural sample)

end digit	decennial age range							
	0—9	10—19	20—29	30—39	40—49	50—59	60—69	70—79
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
0	4.6	3.5	7.9	4.8	7.2	10.6	1.8	
1	−2.0	−2.0	−1.5	−0.1	0.5	−0.4	−0.8	
2	3.5	−0.9	−1.7	−2.3	−0.3	−0.3	−1.2	
3	−2.1	1.8	−1.1	−0.8	−0.5	−0.7	0.6	
4	−0.3	−0.3	−3.3	−0.2	−1.0	0.1	−1.7	
5	−0.5	8.0	2.0	2.7	−3.5	−0.9	1.0	
6	1.0	−2.9	−0.1	−2.1	−1.2	−2.3	0.6	
7	−3.1	−0.4	−0.9	0.4	−1.1	−1.7	0.6	
8	1.5	−1.9	−1.4	−1.6	−0.3	−2.9	−0.6	
9	−2.6	−1.3	0.1	−0.8	0.2	−1.5	−0.3	

5.6. The differences for a number of consecutive end digits immediately before age 60 are all negative and comparatively larger, and the differences generally change sign in this area. The conditions in the immediate neighbourhood of age 25 are even less distinctive. Some age bias is however now suggested for age 16, which appears to have gained at the cost of ages 13–15. But the evidence is far from conclusive; the big and mounting concentrations at the preferred digits are no doubt mainly responsible for this lack of conclusiveness. And specific localised age bias is also relatively milder in India.

Section Six

MEASURES OF CONCENTRATION AND DISTORTION

6.1. It is obvious that the extent of concentration at each digit and the resulting aggregate distortion have to be measured before any efficient set of grouping could be built up. The concentration at each digit can be measured by comparison of deviations of actual numbers returned at individual ages from the expected numbers in the corresponding ages obtained by graduation. It will be assumed however that no graduation of the age returns has been done, as the problem of grouping ceases to exist if a good unbiased graduation, influenced by subjective choice of the operator, is already available; any set of grouping will be equally efficient when built up from such graduated numbers.

6.2. At one time, the total numbers returned at each end digit used to be compared to a tenth of the population, to get a measure of the integer bias and of the total distortion, on the assumption that all the end digits should occur with equal frequency if there was no integer bias. King (1916) then pointed out that the starting integers 0, 1, 2, ... got additional weightage in that order in such comparison[6·1]. The difficulty was resolved by Myers (1940) who used a blended population for the purpose in his index of concentration[6·2]; each digit was put successively at each place from first to tenth in the component populations by Myers, so that they got balanced weight in the resulting blended population. Myers started with age 10 and showed that the average concentration values for the United Kingdom 1911 Census age returns yielded by his method were remarkably close to the values obtained by King by comparison of the numbers returned against the graduated numbers.

6.3. A much simpler measure of concentration was evolved in connection with the analysis of distortion in NSS age returns. In a normal population structure, where the numbers alive gradually fall with age, digit '0' will show up a higher concentration than really attached to it if the total numbers at different end digits of age were all compared to a tenth of the total population starting with age '0'. But this difficulty will be circumvented if for each different end digit of age, the population starting with that particular digit was only taken into account : thus, for digit '0' the proportion of the number returned with end digit '0' ages to the total population aged zero and above, for digit '1' the proportion of the number returned with end digit '1' ages to the population aged one and above, and so on, be taken. The proportions for different digits will not usually add up to unity under this method, but when reduced to a unit base the proportions should give proper relative measure of concentration for each of the individual digits.

6.4. The digit '0' may still get a slight weightage under this method owing to the incidence of the high infant mortality; but that will not be material; and the practical consideration is there that under-reporting of infants, a common feature of censuses and surveys, tends to offset this.

6·1 *Journal of the Institute of Actuaries, Vol. XLIX,* p. 301.

6·2 *Transactions of the Actuarial Society of America, Vol. XLI, Part 2, No. 104,* pp. 402-415.

6.5. The measure of concentration suggested above and the index of aggregate distortion derived from it, are defined below in algebraic symbols. If n_{10v+u} denote the number actually returned at age $10v+u$ then 'm_u' the measure of concentration at digit 'u' is given by

$$m_u = \frac{\sum_{v=0} n_{10v+u}}{T_u} \times 10, \text{ where } \quad T_u = \sum_{x=u}^{9} n_x + \sum_{v=1}^{9} \sum_{u=0} n_{10v+u}.$$

And 'I' the index of aggregate distortion is given by

$$I = \sum_{0}^{9} |m_u - 1|$$

It will be arcle that 'I' will tend to zero in ideal conditions, if there were no distortion.

6.6 The measure of concentration for each digit and the index of aggregate distortion of the NSS medium for rural India, urban India and the city of Calcutta are given in Table (6.1). The measure and index for the population aged 40-above in the rural sector, for males and females separately in the urban sector (where the sex differential was found highest), and for the population of household-income group Rs.100 and below per month in the city of Calcutta have been actually presented in the table, to convey an idea as to how the concentrations vary from one population segment to another. The measures and index for Census 1951 in Uttar Pradesh individual age returns are also shown for comparison.

TABLE (6.1): MEASURES OF CONCENTRATION AT INDIVIDUAL END-DIGITS AND INDEX OF AGGREGATE DISTORTION IN AGE RETURNS

(NSS 4th round 1952, Calcutta Employment Survey 1953, and Census of India 1951)

end digit	NSS 4th round : all India					Calcutta Employment Survey (1,056 households)		Census 1951[6·4]
	rural (28,918 persons)		urban[6·3] (28,715 persons)					
	all ages	aged 40-above	persons	male	female	all income groups	household income ⩽ Rs. 100 per month	U.P. (males)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	1.8	3.1	1.6	1.6	1.7	1.9	2.5	1.9
1	0.6	0.4	0.7	0.7	0.7	0.6	0.6	0.7
2	1.1	0.9	1.1	1.2	1.1	1.2	1.1	1.1
3	0.8	0.5	0.8	0.8	0.8	0.8	0.6	0.7
4	0.8	0.5	0.9	0.9	0.8	0.9	0.7	0.8
5	1.6	2.3	1.4	1.4	1.5	1.5	1.7	1.7
6	0.9	0.6	0.9	0.9	0.9	0.8	0.8	0.9
7	0.7	0.5	0.8	0.8	0.8	0.7	0.6	0.6
8	1.1	0.8	1.1	1.0	1.1	1.1	1.0	1.0
9	0.6	0.4	0.7	0.7	0.6	0.5	0.4	0.6
total	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
index of aggregate distortion	3.2	6.8	2.4	2.4	2.8	3.4	4.6	3.4

6·3 Twelve out of total sixteen part-samples of NSS 4th round were used for this note.
6·4 *Census of India 1951, Paper No. 3, 1954*, pp. 36-37.

6.7. It will be seen that the urban pattern of concentration differed from the rural pattern on the one hand and the city pattern on the other. The most efficient set of grouping for the urban age returns need not therefore be the most efficient for the rural and the city sectors. The fact that not only the pattern of concentration of the household-income group Rs. 100 and below per month of the city of Calcutta was of different nature from the general city pattern, but that the distortion in their age reporting was even greater than in the general rural population, is somewhat unexpected. This apparently results from the break-up of the family and the drift from original community moorings of individuals in the lower income group; and the consequent failure of the applicability of a relative seniority ranking scale within the household and the community. The higher average age of the city population particularly in the lower income level, could also be a contributory cause. The implications of these differentials will be dealt with further in the next section.

6.8. It is easy to see from the distribution of numbers returned at individual ages (the diagrammatic representation of Chart (1) for example) that the force of concentration is comparatively low in the young age range; as could have been anticipated on *a priori* grounds, it gradually increases with advancing age. In the beginning, pulls of even over odd is most effective; then the pulls of '4' and '6' fade out giving place to increased pull for the middle digit '5'. Pulls of '0' and '5' dominate the middle age range, with '0' building up as age advances further. The position is complicated by existence of special pulls of the nature of age bias.

6.9. While the mounting nature of the pull of concentration has been noticed by earlier operators, no relative measures of deviation for the different age ranges appear to have been used so far. The root mean square deviation in decennial age ranges, calculated as the square-root of the sum of the squared deviations between the numbers actually returned and the expected frequencies, can provide such measures[6·5]. The difficulty which perhaps weighed with the operators in this field was that of estimating the expected frequencies. Bnt simple assumptions like linear fall in expected frequencies between quinary pivotal values (estimated from the numbers actually returned by suitable grouping) should serve the purpose. The set of expected frequencies produced by even such crude assumptions would take account of the general shape of the true distribution and thus give satisfactory relative measures. The measures of concentration, taken along with such relative range measures of deviation could only give a proper insight into the extent and spread of the distortion.

[6·5] A number of other alternative measures of deviation are of course possible: one could be $\sum_{u=0}^{9} \frac{(n_{10v+u}-r_{10v+u})^2}{r_{10v+u}}$, giving the χ^2-analogue of the distribution in the decennial age range $10v$ to $10v+9$.

6.10. In algebraic symbols, i_v the relative range measure of deviation of the decennial age range $10v$ to $10v+9$ is defined as,

$$i_v = \frac{\sqrt{\sum_{u=0}^{9} (n_{10v+u} - r_{10v+u})^2}}{\sum_{u=0}^{9} r_{10v+u}}$$

where r_{10v+u} is the expected number at age $10v+u$.

The expected number r_{10v+p} at pivotal age $10v+p$,($p = 4,9$ for the 2 : 7 grouping adopted), was taken as

$$r_{10v+p} = \frac{1}{5} \sum_{u=p-2}^{p+2} r_{10v+u} \text{ ; and the espected numbers at other age,}$$

$$r_{10v+u} = r_{10v+p} - \frac{u-p}{5} (r_{10v+p} - r_{10v+p+5}), \quad \text{where, } u = p+1,\ p+2, \ldots, p+4.$$

6.11. The relative range measures of deviation in the successive decennial age ranges for the population returned at individual ages in NSS for rural sector are given in Table (6.2). The progressive increase in the measures of deviation with advance of age is clearly brought out in columns (2) and (3) of the table. The root mean square deviation per cent of age (obtained by dividing the deviation per individual by the middle age of the range for simplicity and multiplied by 100) are also shown in the table : the interesting fact that the average deviation per unit of age attained is nearly uniform in all age ranges emerges from this part of the table.

TABLE (6.2) : RELATIVE RANGE MEASURES OF DEVIATION IN DECENNIAL AGE RANGES

(NSS 4th round 1952, all-India rural samples : 28,918 persons)

age range	deviation per individual		deviation percent of age	
	male	female	male	female
(1)	(2)	(3)	(4)	(5)
1. 0— 9	0.03	0.02	0.61	0.43
2. 10—19	0.09	0.09	0.56	0.58
3. 20—29	0.15	0.15	0.61	0.59
4. 30—39	0.25	0.19	0.70	0.54
5. 40—49	0.29	0.25	0.63	0.56
6. 50—59	0.30	0.33	0.54	0.60
7. 60—69	0.37	0.50	0.57	0.76
8. 70—79	0.42	0.41	0.57	0.54
9. 80—89	0.36	0.58	0.42	0.69
10. 90—99	0.56	0.48	0.59	0.51

Section Seven

GROUPING EFFICIENCY

7.1. The efficiency of a set of grouping is traditionally assessed by the difference between the sum of the actual concentrations at the end digits comprised in the group and the ideal values. $E_{0:5}$ the efficiency index of the quinary set of grouping 0 : 5 for example is given by $\sum_{0}^{4} m_x - 5$: it is obvious that the complementary value $\sum_{5}^{9} m_x - 5$ will be the same, with just the sign reversed. This method was not altogether satisfactory in that the weight of numbers lay in the age range below 20 in population formations like India, and the group efficiency was accordingly determined to a greater extent by the pattern of concentration in this young age range : the chosen set of groups are however used throughout the span of life, and also for different socio-economic segments of the population, where the patterns of concentration are different.

7.2. It has been seen earlier how the distortions are comparatively small in the lower age range up to 20 and gradually increase with age. The relative efficiency of a set of grouping in the higher age ranges should thus provide a better indicator. It was therefore proper to decide on the most efficient set of grouping from a study of the behaviour of the various possible sets in different age ranges; the study should preferably extend to other socio-economic segments of the population. Table (7.1) gives the efficiency indices of the various possible sets of grouping in some important population segments.

TABLE (7.1) : GROUP EFFICIENCY INDEX OF DIFFERENT SETS OF GROUPING

(NSS 4th round 1952, Calcutta Employment Survey 1953, and Census 1951)

set of grouping	NSS 4th round : all-India					Calcutta Employment Survey (1,056 households)		Census 1951 U.P. (males)
	rural (28,918 persons)			urban (28,715 persons)				
	all ages	aged 30-above	aged 40-above	male	female	all income groups	household income ⩽Rs. 100 p.m.	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. 0 : 5	0.15	0.31	0.43	0.14	0.11	0.35	0.46	0.14
2. 1 : 6	−0.10	−0.36	−0.44	−0.01	−0.16	−0.07	−0.27	−0.05
3. 2 : 7	0.20	−0.06	−0.24	0.21	0.07	0.17	−0.01	0.16
4. 3 : 8	−0.24	−0.52	−0.62	−0.21	−0.27	−0.26	−0.49	−0.26
5. 4 : 9	+0.08	−0.18	−0.31	0 08	0.09	0.02	−0.11	0.01

7.3. Table (7.1) demonstrates how the relative variations in the group efficiencies, spread out in higher age segments of the same population. The urban and Calcutta samples were not analysed further in age segments, but the Calcutta sample was examined in an economic segment; in the household income $\leqslant$ Rs. 100 p.m. population segment the variation again scattered wider. The group efficiency indices of the Census 1951 U.P. individual age distribution (1% sample)[7·1] is also shown in the table for comparison.

7.4. The rural sector is by far the more important; but on the basis of all ages there was not much to choose between the different sets of grouping, though the 4 : 9 and 1 : 6 sets had low indices. It has been pointed out earlier why the all ages index was not satisfactory. The 2 : 7 set shows the definite minimum indices at the higher age segments : it similarly has a definite minimum index in the lower income group, which again was by far the more important economic segment. The real test of efficiency is thus satisfied by the 2 : 7 set for which the index remains more stable and comparatively low in all the different segments, particularly where the indices for the other sets soar up : the index for this set also shows more changes in sign in passing through the population segments, which make for more balanced distribution of the group errors between it and its complementary group.

7.5. The 2 : 7 set of grouping was therefore adopted in analysis, interpretation and presentation of NSS data on age (as well duration). With a general tendency to over-estimate, the ages with the end digits '0' and '5' were likely to draw comparatively more from the ages below and the 2 : 7 set of grouping with the maximum concentration digits '0' and '5' placed towards the end of the groups, was also efficiently constituted to take account of this aspect of the estimation error.

7.6. Though the question of the most efficient set of grouping for the Indian Census was not in issue, examination and some discussion of the efficiency of grouping in the Census medium became inevitable. The mediums of collection of information and the types of errors are different for the Census and the NSS ; the population is, however, the same and the relative efficiencies of the various sets of grouping could at least be expected to remain undisturbed as between them.

7.7. In Census 1931 Report[7·2] the 2 : 7 grouping was recommended after analysis of the age data in individual years on traditional lines. No detailed examination was done of Census 1941 age data. Numbers returned at individual ages in Census 1951 were not available for all India. Analysis of concentration and of group efficiency of the Census age material for Uttar Pradesh (U.P.), the only State which constitutes a Census population zone by itself, is shown in Table (6.1) and (7.1) : in Census 1951 Age Tables[7·3] some detailed examination of the age data of the same

---

7·1 *Census of India 1951, Paper No. 3, 1954*, pp. 36-37.

7·2 *Census of India 1931, Vol. I, Part I*, p. 135.

7·3 *Census of India 1951, Paper No. 3, 1954*, p. 22.

State was done to decide about the most efficient set of grouping. In the Age Tables the 2 : 7 set has been described as the standard grouping, but the 3 : 8 set has been recommended in the same breath as the 'proper' set; the belief that the dominant digits '0' and '5' should draw nearly equally from either side apparently tipped the scales in favour of the 3 : 8 set.

7.8. Primary grouping of the Census age returns in the 3 : 8 set however produced a saw-tooth distribution and a roller-type formula had to be applied to these quinary group values to get a smooth run. This was achieved by taking the weighted average of the group frequency concerned and the two adjoining group frequencies. The smooth set of group frequencies ultimately operated on for graduation purposes in Census 1951 thus rested on the assumption that the dominant digits drew from 7 individual ages on either side : such assumption looks stretched on the face of it. Table (7.1) showed clearly how the 3 : 8 set is the least efficient for the Indian situation. The problem of grouping efficiency exists even behind the operation of graduation : good graduation can only flow from an efficient set of group pivotal values.

7.9. It is interesting to note that if the numbers returned in individual ages in Census 1951 are grouped under the 2 : 7 set, the total deviations of the actual group frequencies from the expected (built up from the corresponding graduated individual frequencies) are smaller than similar total deviations of actual from expected under the 3 : 8 set of grouping; this was actually verified for the individual age distributions of Uttar Pradesh and Madras, special notice of which was taken in the Census 1951 Age Tables[7·4] to select the 'proper' grouping. When it is realised that the graduation itself is based on the 3 : 8 set, greater confidence is gained about the superiority of the 2 : 7 set.

7.10. Table (7.2) gives the comparative deviations of the actual numbers returned from the expected for the two competing sets of grouping 2 : 7 and 3 : 8 for the Census 1951 (males) population of Uttar Pradesh and Madras[7·4].

7.11. The problem of grouping has been dealt so far with reference to age returns in individual years. But there may be situations when collection in individual years is either not possible or advised. The considerations guiding the selection of the most efficient groups will be altogether different if ages are as a rule not known or cannot be estimated in individual years. An ingenious suggestion made by R. Bachi (1951) to meet a somewhat similar situation deserves special mention in this context. He advised that all series which could not be collected by individual years might be collected for individual end digits '0' and '5' only, and for part-groups of end digits 1–4 and 6–9, as dominant distortions will be disclosed thereby[7·5]. Bachi further suggests routine methods of allocating the numbers returned at each of the preferred end digits '0' and '5', to the two bordering part-groups :

[7·4] *Census of India 1951, Paper No. 3, 1954,* pp. 36-37 and 68-69.

[7·5] *Bulletin of the International Statistical Institute, 1951, Vol. XXXIII, Part IV,* pp. 218-221.

allocation in equal parts, or in proportion to the weights of the part-groups, or in inverse proportion to the relative broad aggregate measures of concentration of the part-groups (with adjustments for declining numbers with age) are the alternatives proposed.

TABLE (7.2): COMPARATIVE DEVIATIONS BETWEEN CENSUS NUMBERS RETURNED AND EXPECTED UNDER DIFFERENT SETS OF GROUPING

	grouping set 2 : 7					grouping set 3 : 8				
	age group (years)	number returned (000)	number expected (000)	deviation (2)—(3) (000)		age group (years)	number returned (000)	number expected (000)	deviation (2)—(3) (000)	
	(1)	(2)	(3)	(4)		(1)	(2)	(3)	(4)	
					Uttar Pradesh (males) : Census 1951					
				+	−				+	−
1.	2— 6	4287	4191	96		3— 7	4244	4241	3	
2.	7—11	4129	4065	64		8—12	4453	3971	482	
3.	12—16	3821	3606	215		13—17	3151	3505		354
4.	17—21	2744	3154		410	18—22	2981	3071		90
5.	22—26	2962	2806	156		23—27	2631	2745		114
6.	27—31	2476	2551		75	28—32	2709	2497	212	
7.	32—36	2417	2293	124		33—37	2075	2238		163
8.	37—41	1986	2026		40	38—42	2114	1968	146	
9.	42—46	1737	1764		27	43—47	1553	1710		157
10.	47—51	1538	1490	48		48—52	1609	1427	182	
11.	52—56	1084	1182		98	53—57	961	1117		156
12.	57—61	908	871	37		58—62	922	809	113	
13.	62—66	492	582		90	63—67	426	530		104
14.	total	30581	30581	740	740	total	29829	29829	1138	1138
average percentage deivation					4.84	average percentage deviation				7.63
					Madras (males) : Census 1951					
				+	−				+	−
1.	2— 6	3701	3644	57		3— 7	3607	3621		14
2.	7—11	3279	3398		119	8—12	3618	3339	279	
3.	12—16	3541	3155	386		13—17	2949	3093		144
4.	17—21	2500	2806		306	18—22	2613	2724		111
5.	22—26	2500	2433	67		23—27	2326	2366		40
6.	27—31	2172	2165	7		28—32	2233	2118	115	
7.	32—36	1951	1968		17	33—37	1793	1927		134
8.	37—41	1823	1774	49		38—42	1885	1730	155	
9.	42—46	1482	1555		73	43—47	1360	1506		146
10.	47—51	1405	1320	85		48—52	1429	1269	160	
11.	52—56	946	1062		116	53—57	857	1007		150
12.	57—61	862	804	58		58—62	871	751	120	
13.	62—66	470	548		78	63—67	409	499		90
14.	total	26632	26632	709	709	total	25950	25950	829	829
average percentage deviation					5.32	average percentage deviation			6.39	

7.12. But the alternative allocations suggested by Bachi did not take account of the bias to over-estimate. Collection of age returns in individual years is possible in the situation contemplated by him, and small sample analysis does not involve much extra cost or time. A sample of the population (or of the Census slips) can indicate the estimation and other biases and should yield sufficiently accurate estimates of concentration and group efficiency. The most efficient set of grouping could be determined in this manner, in advance of the general tabulation, which may be done straightaway after that in the set of grouping adjudged most efficient.

APPENDIX 0

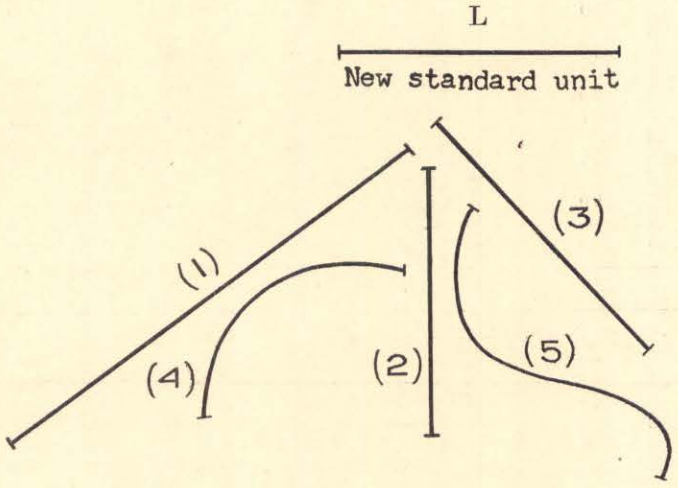
PROFORMA OF SCHEDULE USED IN THE ESTIMATES AND ESP STUDY

ESTIMATION AND ESP STUDY

The Demography Unit will be very much obliged if you kindly fill up the particulars below in your spare time and return the sheet to the Unit (Sri Samarendra Nath Mitra) early.

Roll No..... Date.....

I. In terms of 'L' a new unit of linear measurement specified below, eye-estimate the lengths of the following five lines to two places of decimals and record the estimates :



	Estimated length (0.00)
(1)	
(2)	
(3)	
(4)	
(5)	

II. The middle 3 digits of the following seven digited numbers are missing : please complete the numbers by filling up the middle blank space with the digits that you think, on your first guess, might have been there.

(1)	93	85
(2)	27	47
(3)	45	90
(4)	12	08
(5)	66	19

Thank you !

19 6 54

(Ajit Das Gupta)

APPENDIX 1

DETAILED TABLES

TABLE 1(1): DISTRIBUTION OF INDIVIDUALS IN AGE-ASSESSED MINUS AGE-STATED CLASSES BY RELIGION

(NSS, WBSD Study 1954)

age-assessed minus age-stated	city (170 households)			other urban (405 households)			rural (754 households)		
	total	Hindu	others	total	Hindu	others	total	Hindu	others
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. age-assessed = age-stated	495	422	73	1552	1296	256	2733	2160	573
(%)	(83.9)	(86.5)	(71.6)	(86.2)	(87.0)	(82.3)	(77.6)	(80.4)	(68.5)
2. age-assessed < age-stated	20	15	5	70	53	17	274	184	90
(%)	(3.4)	(3.1)	(4.9)	(3.9)	(3.6)	(5.5)	(7.8)	(6.9)	(10.8)
3. age-assessed > age-stated	75	51	24	178	140	38	515	342	173
(%)	(12.7)	(10.4)	(23.5)	(9.9)	(9.4)	(12.2)	(14.6)	(12.7)	(20.7)
4. total	590	488	102	1800	1489	311	3522	2686	836
(%)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

TABLE 1(2): DISTRIBUTION OF INDIVIDUALS IN AGE-ASSESSED MINUS AGE-STATED CLASSES BY MOTHER TONGUE

(NSS, WBSD Study 1954)

age-assessed minus age-stated	city (170 households)			other urban (405 households)			rural (754 households)		
	Bengali	Hindi	others	Bengali	Hindi	others	Bengali	Hindi	others
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. age-assessed = age-stated	370	72	53	1120	261	171	2598	28	107
(%)	(84.9)	(75.8)	(89.8)	(91.2)	(68.9)	(88.6)	(78.1)	(70.0)	(68.1)
2. age-assessed < age-stated	15	4	1	38	27	5	248	5	21
(%)	(3.4)	(4.2)	(1.7)	(3.1)	(7.1)	(2.6)	(7.5)	(12.5)	(13.4)
3. age-assessed > age-stated	51	19	5	70	91	17	479	7	29
(%)	(11.7)	(20.0)	(8.5)	(5.7)	(24.0)	(8.8)	(14.4)	(17.5)	(18.5)
4. total	436	95	59	1228	379	193	3325	40	157
(%)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

TABLE 1(3): FREQUENCY DISTRIBUTION OF THE NUMBER RETURNED AT EACH INDIVIDUAL AGE BY SEX

(NSS 4th round 1952, all-India urban sample)

age (years)	male	female	persons	age (years)	male	female	persons
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
0	394	372	766	45	319	252	571
1	368	389	757	46	77	65	142
2	460	389	849	47	71	52	123
3	378	375	753	48	128	101	229
4	391	348	739	49	75	47	122
5	361	350	711	50	250	327	577
6	350	359	709	51	54	38	92
7	346	337	683	52	132	85	217
8	349	367	716	53	49	49	98
9	337	289	626	54	64	38	102
10	393	395	788	55	170	189	359
11	283	287	570	56	60	70	130
12	436	416	852	57	48	37	85
13	283	263	546	58	53	48	101
14	358	326	684	59	25	17	42
15	297	283	580	60	203	207	410
16	353	319	672	61	20	31	51
17	246	241	487	62	41	44	85
18	391	396	787	63	23	18	41
19	222	201	423	64	28	21	49
20	399	445	844	65	84	110	194
21	223	161	384	66	19	15	34
22	353	349	702	67	27	20	47
23	224	149	373	68	16	22	38
24	261	216	477	69	12	8	20
25	401	418	819	70	80	100	180
26	261	227	488	71	7	8	15
27	195	168	363	72	17	12	29
28	288	254	542	73	11	15	26
29	118	123	241	74	9	5	14
30	487	429	916	75	28	40	68
31	123	98	221	76	2	7	9
32	276	202	478	77	6	1	7
33	135	101	236	78	4	5	9
34	150	95	245	79	1	5	6
35	410	332	742	80	20	39	59
36	155	155	310	81	4	2	6
37	115	108	223	82	6	7	13
38	174	149	323	83	–	1	1
39	88	64	152	84	1	6	7
40	378	378	756	85	5	6	11
41	82	51	133	86	–	1	1
42	149	121	270	87	–	–	–
43	86	73	159	88	1	–	1
44	81	88	169	89	2	–	2