MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Konrad-Zuse-Strasse 1 · D-18057 Rostock · Germany · Tel +49 (0) 3 81 20 81 - 0 · Fax +49 (0) 3 81 20 81 - 202 · www.demogr.mpg.de

MPIDR Working Paper WP 2024-001 l January 2024
https://doi.org/10.4054/MPIDR-WP-2024-001

New adjustment procedure for distortion in age distribution

Afza Rasul
Jamal Abdul Nasir
Dmitri A. Jdanov l jdanov@demogr.mpg.de

New adjustment procedure for distortion in age distribution

Authors

Afza Rasul[1,2,3], Jamal Abdul Nasir[2], Dmitri A. Jdanov[3]

[1]Department of Statistics, Lahore College for Women University, Lahore-Pakistan
[2]Department of Statistics, Government College University, Lahore-Pakistan
[3]Laboratory of Demographic Data, Max Planck Institute for Demographic Research, Rostock-Germany

Abstract

Accurate age data is a prerequisite for any demographic inquiry. Unfortunately, in many developing countries visible age heaping is present in census and survey data of reported age at the time of census or survey.

In this article, a new method is proposed for age adjustment of the respondent current age at the time of interview/data collection. The method is based on the rectangular distribution probabilities for terminal digits of age. The algorithms-based method is used to estimate true/adjusted age distribution in the presence of age heaping/age misreporting.

Application of the method is performed on the most recent demographic and health survey data from Afghanistan, Bangladesh, Pakistan, India, Ethiopia, and Gambia.

UN Criteria for age accuracy is used to check the accuracy of adjusted/true age distribution. The result revealed that after adjustment of the terminal digit by the proposed method of digit shift the adjusted age distributions are perfectly accurate. The method will be applicable to survey and census data. The method will be very useful in fertility analysis where the individual year of age of women plays an important role.

Keywords: Whipple Index, Age Heaping, Age misreporting, Digit preferences, Digit avoidance, Adjusted age

Background

Age is the most important study variable in demographic research. By definition, the current age is the number of completed years by any given moment by which one is telling his/her age. For example, if an individual reports his/her age as 38 years, he/she is currently somewhere between 38 and 39. In socio-demographic research and surveys, misrepresentation of age results in wrong statistical estimates, hence, can mislead policy stakeholders in formulating effective policies.

Age statements are commonly affected by two types of errors: age heaping, which is the tendency to round ages to specified digits (0 or 5), and systematic exaggeration or underestimation of ages. Age heaping, also known as age preference or digit preference, mostly occurs when people do not know their true ages and report them in round numbers or urge the enumerators to write down whatever age they believe is appropriate. Researchers' findings showed that in many DHS surveys, age heaping, and digit preference are present (Fayehun et al., 2020; Randall & Coast, 2016; Singh et al., 2022; Szołtysek et al., 2018). Age misreporting in other surveys and censuses in the developing world is also very common (Pardeshi, 2010; Pullum, 2005, 2006; Samuel, 2018; Singh, 2017; Szołtysek et al., 2018), and estimates drawn from misreported age, results in uncertainties of age distributions.

Data adjustments are needed for various official and non-official policy stakeholders. Quantitative estimates of vital events were observed to be lacking the actual level of vital events particularly mortality and fertility estimates (Caldwell, 1966). Age distortion seemed substantially impacting on policy needed vital estimates (Krafft et al., 2021; Machiyama, 2010).

The debate about age misreporting among demographers is not new. A century ago, George Chandler Whipple (1866-1924), an American demographer give an index to measure the tendency for human age misreporting. Later several techniques have been developed and used to measure age misreporting in surveys and censuses for age distributions. Myers' index (Myers, 1940, 1954), Bachi's index (Bachi, 1951), Carrier index (Carrier, 1959), and Ramachandran index (Ramachandran, 1965) have been developed and used to check the quality of age data. Some modified versions of the Whipple index were proposed: Modified Whipple index (Noumbissi, 1992), total modified Whipple index (Spoorenberg & Dutreuilh, 2007), Whipple-type index or Whipple 3 index (Poston & Micklin, 2005; Poston et al., 2003; Poston Jr et al., 2000), ABCC index (A'Hearn et al., 2009) and remodified Whipple Index (Nasir & Hinde, 2014).

All indices described above identify age misreporting, age heaping, or age distortion. However, there are a few remedies that are used to treat this misreporting, heaping, and distortion. In the last decades of the previous century, some earlier attempts have been made by demographers to adjust age misreporting (Bhat, 1990; Demeny & Shorter, 1968; Gupta, 1975; Ntozi, 1978) from census data, but all these methods have a rare application due to their limitations, underlying assumptions and lack of census practices in some developing countries. Demeny-Shorter suggested a technique to adjust age by combining age data from two censuses while analyzing Turkish census data (Demeny & Shorter, 1968). The mathematical formulation of the Demeny-Shorter method has been described along with a critical evaluation in the articles (Gupta, 1975; Ntozi, 1978). Gupta (1975) identified that the Demeny-Shorter technique made an implicit assumption that the two censuses had equal age patterns so this method fails in case of failure of assumptions. To solve this problem, Gupta (1975) proposed some "more general method" for the correction of age misreporting in the census data but as the age gap between the age structures widens, the approach faces an increasing difficulty of the inconsistency of the results with the underlying assumptions. Therefore (Ntozi, 1978) developed a new technique based on the same concept as the Demeny-Shorter method, but using age data from three consecutive censuses rather than two. This approach was applied to data from Turkey's censuses and has shown to be superior to the Demeny-Shorter method in some circumstances. Unfortunately, due to a lack of the necessary series of censuses, the three-census approach cannot be applied to data from most developing countries.

Different smoothing techniques are also applied to smooth the age distribution in the presence of heaping or distortion and are widely used in literature (Siegel & Swanson, 2004; Yusuf et al., 2014). The simplest way to smooth age data is the use of moving average methods, however, this method has the limitation that a certain number of ages in the beginning and at the end vanish. Another simple method widely used in literature is the aggregation/grouping of age data. Grouping of age variable in 5-years and 10-years age groups is assumed to be a very useful technique to smooth age distribution; however, in some demographic analyses, such as fertility analysis, age grouping does not yield beneficial findings. There are many smoothing formulas used to smooth age data combined in 5 years of age groups; Carrier and Farrag (Carrier & Farrag, 1959), Karup–King–Newton (Carrier & Farrag, 1959) , Arriaga's formula , UN method (UnitedNations, 1952),the strong smoothing method (Arriaga, 1968) or moving average method, and others. All of them have limitations. Some of the methods assume that the total reported population in the census is correct and only the age distribution is wrong. Carrier and

Farrag, Karup–King–Newton, and Arriaga, assume that the population total for 10 years age group is correct while the UN method and strong method are used to adjust the heaping errors in successive intervals.

Recent works (A'Hearn et al., 2009; Nasir & Hinde, 2014; Spoorenberg & Dutreuilh, 2007) focused more on figuring out the pattern of age distortion in survey and census data. There has been less focus on correcting these age distortions which has been proven a big concern by using the numerous methods cited above.

Enormous literature exists to explore the issue of age misreporting though various indexes, Whipple index including its modifications, Myer's, Bachi's are some well-known indexes. The indexes values indicate only the quality of age reporting error. In addition, the preferences or avoidance of certain terminal digits between 0 and 9 inclusive over the other may well explained by index values. Alternatively, statistical models in particular logistic regression model identify the preference or avoidance of terminal digit. Both indexes and statistical model are seemed to be incapable to construct and adjusted age distribution. In this paper, we propose a new method to adjust distortions of age distributions using various survey data sets.

Data and Methods

Data: The present research uses data from internationally representative household surveys, namely Demographic and Health Surveys (DHSs) (ICF, 1985-2023). The DHS program has been collecting accurate and representative data on population, health, HIV, and Nutrition through more than 400 surveys in over 90 countries. We use the most recent waves of the standard DHS surveys. These surveys have large sample sizes and are typically conducted about every five years. These DHSs are cross-sectional surveys that use several different sets of questionnaires. One of the questionnaires is the household questionnaire, used to collect information for all household members. Based on this household questionnaire survey, quality of age data from individual participants aged 23-62 years has been undertaken.

Methods: We use Whipple index to identify most problematic data series and to check the quality of adjusted data series. The proposed new method is further development of the method of Digit Shifts by Nasir (2013) which is based on Multinomial Regression Model for Terminal Digits.

Whipple index. Among the indices to identify incorrect reporting of age distribution Whipple Index (*WI*) (Siegel & Swanson, 2004). Whipple Index is based on the rectangular distribution property and can be calculated as follows:

$$WI = \frac{(f_{25}+f_{30}+f_{35}+\cdots+f_{60})}{\frac{1}{5}(f_{23}+f_{24}+f_{25}+\cdots+f_{62})} * 100 , \qquad (1)$$

where f_i is the total number of persons with reported age *i*. The value of the *WI* in any population with no large changes in fertility, mortality, and migration for a reported period of study/survey/ census would be 100. The United Nations (UN) recommended that if Whipple's index deviates by less than 5 percent from a perfect standard then we consider age to be reported accurately (UnitedNations, 1955). The standard recommended by the UN is as follows;

Table 1. UN standard for quality of age distribution using Whipple index Value

Whipple index Value	Deviation from perfection	Quality of data
<105	<5%	Perfectly Accurate
105-110	5-9.99%	fairly Accurate
110-125	10-24.99%	Moderate
125-175	25-74.99%	Poor/rough
>175	≥75%	Very poor/rough

Multinomial Regression Model for Terminal Digits

In a numerical system generally, age is the combination of two digits (tens and units). Taking units i.e. terminal digit (*0, 1, 2, …,9*) of age as an outcome variable of a multinomial regression model. Let $X_1, X_2, X_3, \ldots, X_k$ is a collection of *k* independent variables (covariates or predictors) with an estimated vector of coefficient as $\underline{\boldsymbol{\beta}} = (\beta_1, \beta_2, \beta_3, \ldots, \beta_K)$. Let $\underline{\boldsymbol{\pi}} = (\pi_0, \pi_2, \pi_3, \ldots, \pi_9)$ be the vectors of probabilities of the digit preference or avoidance. The odds of the digit preference(avoidance) of the multinomial regression model can be:

$$\frac{\pi_i}{\pi_0} = exp\left(\alpha_i + X\underline{\beta_i}\right) \qquad (2)$$

Here α_i is the intercept term, the log odds of the model would be.

$$Log\left(\frac{\pi_i}{\pi_0}\right) = \alpha_i + \sum_{j=1}^{k} \underline{\beta_i}\, X_j \qquad (3)$$

The estimated probabilities $(p_0, p_1, p_2, \ldots, p_9)$ for digit preference/ avoidance can be calculated as:

$$p_0 = \left[\frac{1}{1+E}\right] \text{ and } p_i = \left[\frac{exp\left(\alpha_i + X\underline{\beta_i}\right)}{1+E}\right]$$

Where *i = 1, 2, 3, …,9 and*

$$E = \left\{exp\left(\alpha_1 + X\underline{\beta_1}\right) + exp\left(\alpha_2 + X\underline{\beta_2}\right) + exp\left(\alpha_3 + X\underline{\beta_3}\right) + exp\left(\alpha_4 + X\underline{\beta_4}\right) + exp\left(\alpha_5 + X\underline{\beta_5}\right) + exp\left(\alpha_6 + X\underline{\beta_6}\right) + exp\left(\alpha_7 + X\underline{\beta_7}\right) + exp\left(\alpha_8 + X\underline{\beta_8}\right) + exp\left(\alpha_9 + X\underline{\beta_9}\right)\right\}$$

Here terminal digit *"0"* is taken as the reference category. It is important to note that the estimated probabilities remained the same when we change the reference category from *"0"* to any other (*1, 2, 3, …,9*) terminal digit. A detailed description of multinomial regression can be found in many textbooks, e.g. (Hosmer Jr et al., 2013; Kleinbaum & Klein, 2010).

Method of Digit Shifts

The method of Digit Shift was proposed by Nasir (2013). It was used to estimate the true age distribution of the women respondent of the reproductive age group. The method consists from the following four steps:

Step 1. Estimate an imaginary number of women at the terminal digit. Nasir used the estimated probabilities from a multinomial logistic regression model without covariates for unit digit preference or avoidance. Let p_i be the estimated probabilities of terminal digits i = 0,1, 2, 3, … 9. Then the total number of women reporting each terminal digit (W_i^R) using the estimated probabilities can be calculated as:

$$W_i^R = p_i \times N$$

Where N is the size of the cohort/number of respondents.

Step 2. In the second step matrix (10*10) of the digit *"D"* is constructed. The sum of the all elements in matrix D is the constant (10,000) number of women assuming their terminal digits and the rows sum represents the total number of women reporting each digit estimated in step 1.

$$D = \begin{bmatrix} d_{00} & d_{01} & \cdots & d_{09} \\ \vdots & \vdots & \ddots & \vdots \\ d_{90} & d_{91} & \dots & d_{99} \end{bmatrix}$$

$$D = \left[d_{ij}\right]$$

In matrix D, d_{ij} indicates the assigned number of women reporting digit *i* but the true digit *j (i, j = 0, 1, 2, … 9)*. The algorithm for assigning d_{ij} is based on shifting the least digit shifts. Here, *i* (rows) refer to the terminal digit reported, while j refers to the true terminal digit of the women's age.

$\sum_{j=0}^{9} d_{ij} = W_j^R$ for $i = 0, 1, 2, \dots 9$

Step 3. In this step, the matrix of digit weights is constructed. Let A be the matrix of digit weights than

$$A = \left[\alpha_{ij}\right]$$

Where $\alpha_{ij} = \frac{d_{ij}}{W_i^R}$

And the sum of each row of matrix A is 1.0. i.e

$\sum_{j=0}^{9} \alpha_{ij} = 1$ for any $i = 0, 1, 2, \dots 9$

Step 4. In the last step matrix of true age distribution is generated. The digit weights α_{ij} were used to the observed reported age distribution to find the matrix of true age distribution $\left(W_{ij}\right)$ with the elements

$$W_{xy} = \alpha_{xy} W_x^o$$

Where W_x^o is the observed number of women at each age and W_{xy} is the number of women who reported age x but had true age y, and α_{xy} is the probability that a woman with true age y while reporting her age as x. The matrix of the true women age distribution is written as:

$$\begin{bmatrix} \alpha_{15,15} W_{15}^O & \alpha_{15,16} W_{15}^O & \cdots & \alpha_{15,49} W_{15}^O \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{49,15} W_{49}^O & \alpha_{49,16} W_{49}^O & \dots & \alpha_{49,49} W_{49}^O \end{bmatrix}$$

Finally, the true adjusted distribution of women's age is obtained by summing each column of $(35 \times 35 + \delta)$ matrix; with 35 range of data (15 to 49 years of age) and δ is the arbitrary shifting year. Mathematically Nasir (2013) described the following expression

$$W_y^T = \sum_{k=15}^{49} \alpha_{xy} W_x^O$$

And finally

$$\sum_{k=15-l}^{49+u} W_y^T = \sum_{k=15}^{49} \alpha_{xy} W_x^O$$

Revised Model for Digit Shift [Proposed Model]

Nasir's (2013) model for digit shift is based on the probabilities from multinomial regression with no covariates factor. Moreover, the Nasir model is not based on a rectangular distribution assumption. Woman's age range from 15-49 years is used. Here we use the probabilities based on the rectangular distribution assumption from the Whipple index. Probabilities for each terminal digit are calculated using the Whipple index original formula for each terminal digit of age. We use the age limit of 23 to 63 years which is arbitrary and can be changed to any rectangular distribution of age with equal probability of each terminal digit 0, 1, 2, …, 9; e.g. 15-64 or 20-79 etc. The method is based on the probabilities of terminal digits of the reported age at the time of the data collection.

Step1. Let $\widehat{P_i}$ be the estimated probabilities at each terminal digit $i = 0, 1, 2, \ldots 9$. Then the total number of arbitrary individuals reporting each terminal digit (I_i^R) using the estimated probabilities can be calculated as:

$$I_i^R = \widehat{P_i} \times N$$

Where N is prefixed as 10000 (N can take values as 100, 1000, or 10,000).

[Probabilities of digit preference/avoidance can be obtained by using other approaches like; rectangular distribution, Whipple, Whipple type, modified Whipple, and further modified Whipple indices; however, only the multinomial regression model allows us to check the effect of covariates on age under or over reporting]

Step 2. In the second step matrix (10*10) of the terminal digit *"T"* is constructed. This is the same as described by Nasir (2013). The sum of each column of matrix TD is the constant (1000) number of individuals assuming their terminal digits and the rows sum represents the total number of individuals reporting each terminal digit estimated in step 1.

$$T = \begin{bmatrix} t_{00} & t_{01} & \cdots & t_{09} \\ \vdots & \vdots & \ddots & \vdots \\ t_{90} & t_{91} & \ldots & t_{99} \end{bmatrix}$$

$$T = [t_{ij}]$$

In matrix T, t_{ij} indicates the assigned number of individuals reporting terminal digit *i* but the true terminal digit *j (i, j = 0, 1, 2, … 9)*. The algorithm for assigning t_{ij} is based on shifting the least digit shifts. Here, *i* (rows) refer to the terminal digit reported, while j refers to the true terminal digit of the individuals' age.

$\sum_{j=0}^{9} t_{ij} = I_j^R$ for *i = 0, 1, 2, … 9*

Step 3: In this step, the matrix of digit weights is constructed. Let W be the matrix of digit weights than

$$W = [w_{ij}]$$

Where $w_{ij} = \frac{t_{ij}}{I_i^R}$

And the sum of each row of matrix A is 1.0. i.e

$\sum_{j=0}^{9} w_{ij} = 1$ for *j = 0, 1, 2, … 9*

Step 4: In the last step matrix of true age distribution is generated. The digit weights w_{ij} were used to the observed reported age distribution to find the matrix of true age distribution (I_{ij}) with the elements

$$I_{xy} = w_{xy} I_x^o$$

Where I_x^o is the observed number of Individuals at each age x and I_{xy} is the number of Individuals who reported age x but have true age y, and w_{xy} is the probability that an individual with true age y while reporting his/her age as x. The matrix of the true/adjusted age distribution is written as:

$$\begin{bmatrix} w_{23,23} I_{23}^O & w_{23,24} I_{23}^O & \cdots & w_{23,62} I_{23}^O \\ \vdots & \vdots & \ddots & \vdots \\ w_{62,23} I_{62}^O & w_{62,24} I_{62}^O & \dots & w_{62,62} I_{62}^O \end{bmatrix}$$

Finally, the true adjusted distribution of individuals' age is obtained by summing each column of $(40 \times 40 + \delta)$ matrix; with 40 range of data (23 to 62 years of age) and δ is the arbitrary shifting year. Mathematically

$$I_y^T = \sum_{k=23}^{62} w_{xy} I_x^O$$

Where I^T and I^O are True and observed age of an individual at the time of interview/data collection for census or survey.

And finally

$$\sum_{k=23-l}^{62+u} I_y^T = \sum_{k=23}^{62} w_{xy} I_x^O$$

Here age limit is arbitrary. Can be changed according to the data collected, there Generally

$$\sum_{k=a-l}^{b+u} I_y^T = \sum_{k=a}^{b} w_{xy} I_x^O$$

Where "a" and "b" are the youngest and oldest reported ages of individuals including in survey/census with terminal digits following rectangular distribution.

Algorithm for assigning imaginary respondents at true digits

For any terminal digit *i*, (*i = 0, 1, 2,…,9*) let us assume that *d* is the distance from neighboring digits, then all possible adjacent digits would be

i - d, i + d for *d = 1, 2, 3, 4, 5.*

The first least possible distance would be at *d = 1,* and the next least possible distance would be at *d = 2,* and so on till *d = 5.*

In the first step we will identify the digit preference (i^P) and digit avoidance (i^A) using the criteria:

$$i = \begin{cases} i^P & if \quad \hat{p}_i > 0.10 \\ i^A & if \quad \hat{p}_i < 0.10 \end{cases}$$

Where $\hat{p}_i$ are estimated probabilities using multinomial logistic regression with no covariates [or rectangular distribution or Whipple probability for each terminal digit]. List the digits from most preferred to most avoided digits by using these probabilities. Rank the avoided digits (i_r^A) from the most avoided to the least avoided digit. Find out the least distance digit shift [*i – d, i + d*] for all possible shifts of avoided digits. Construct the matrix of digits by filling in the main diagonal values. Start from the most avoided digit i_r^A, using the least digit shift, adjust it. [For adjustment we will assume that under-reported/avoided terminal digits will start taking observation from immediate neighboring digits, both before or after which are over-reported.

This process will continue till the under-reported digit attain the probability p_i = *0.10*.]. Proceed to adjust all remaining avoided digits. Construct the matrix of digit weights A as;

$$W = [w_{ij}] \quad \text{where} \;\; w_{ij} = \frac{d_{ij}}{Y_j^R}$$

Such that the sum of each row of the matrix is 1.0;

$$\sum_{j=0}^{9} w_{ij} = 1.0 \; \forall_j$$

Finally, find the true digit age distribution using

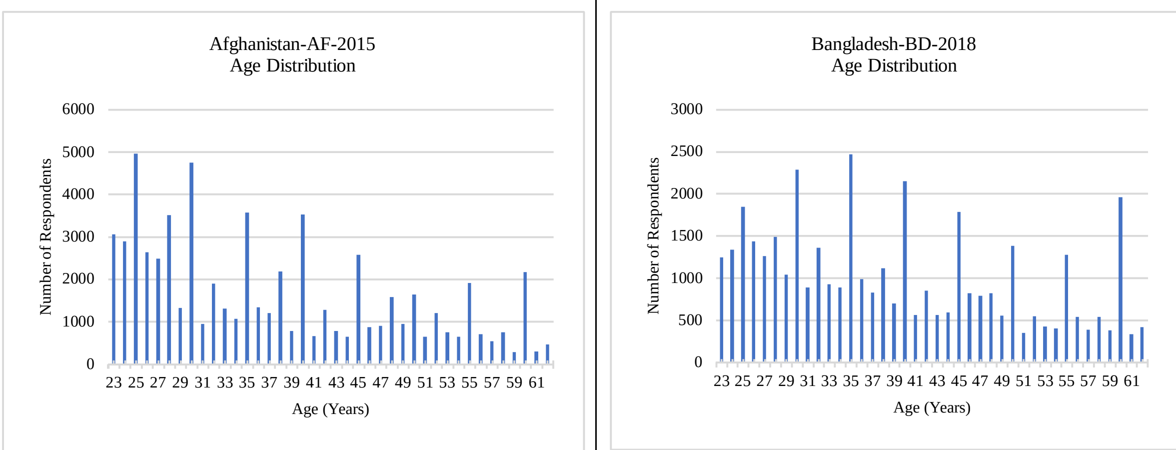
$$I_y^T = \sum_{k=23}^{62} w_{xy} I_x^O$$

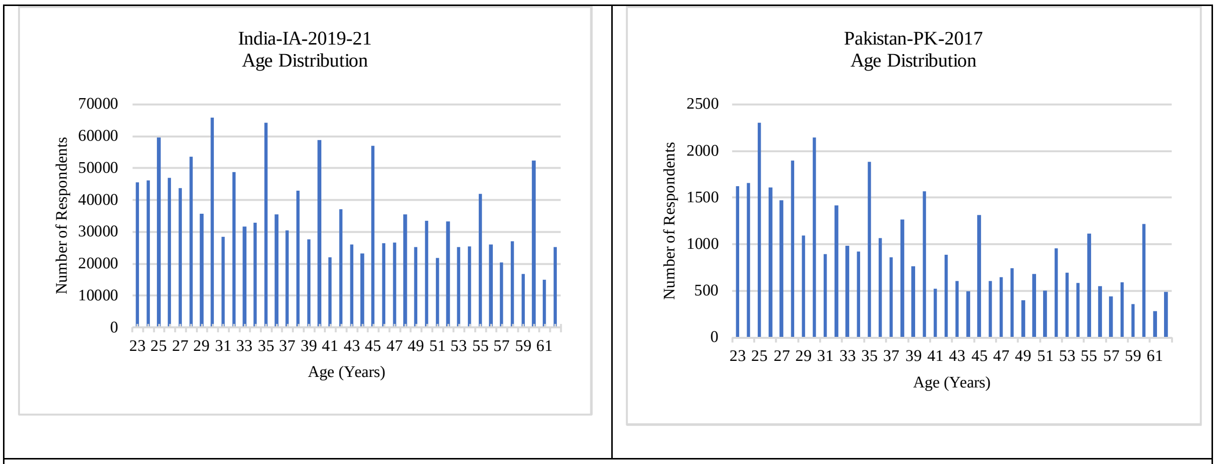
Where I_x^O is the observed number of respondents at each age.

Analysis and Results

A simple graphical presentation is an efficient way to see age heaping at some preferred years of age reported by individuals at the time of interview in DHS surveys at the time of data collection. We take the age range from 23 to 62 years of age as this age range is considered more mature and reliable to report their true age correctly. To express the data quality numerically, the original Whipple Index is used to check the accuracy of the age distribution from Demographic and health survey data sets from all participating countries in DHS surveys. We use the UN recommendation to find age data accuracy based on Whipple Index. Whipple index Value is calculated for all most recent standard DHS data sets (surveys conducted after 2010) for all countries (Table 2).

Figure 1a. Graphical overview of age heaping In DHS surveys for selective Asian countries

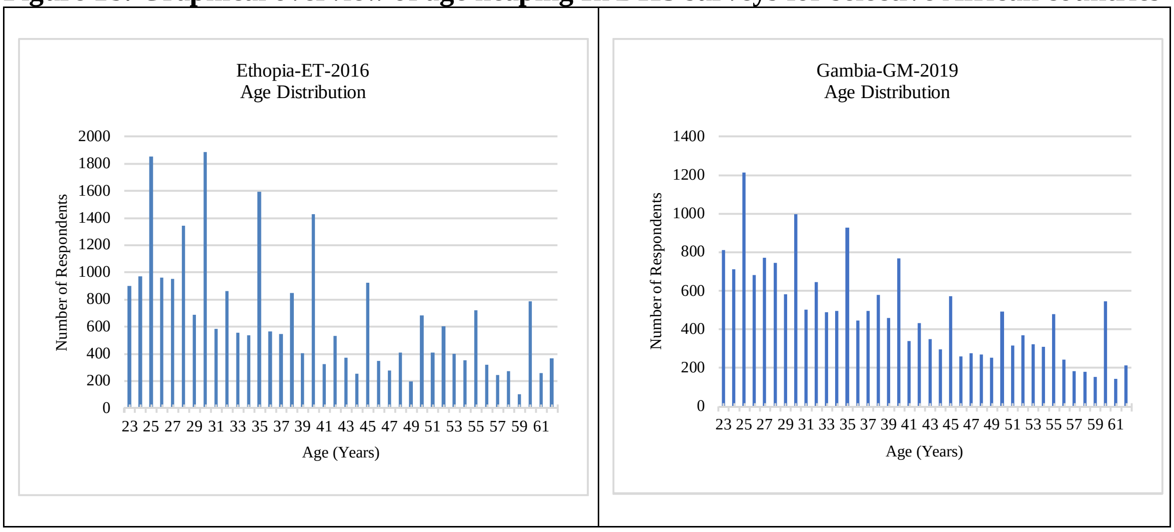


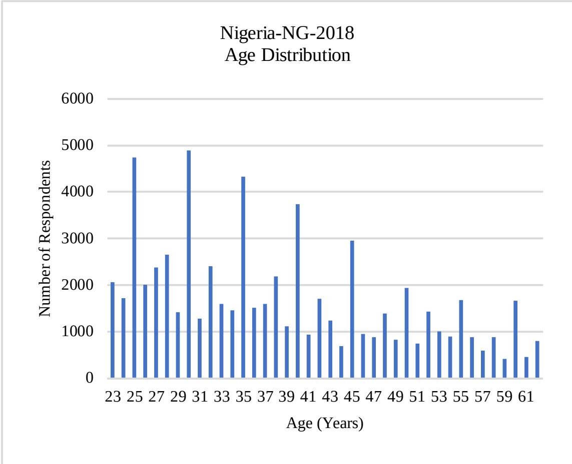


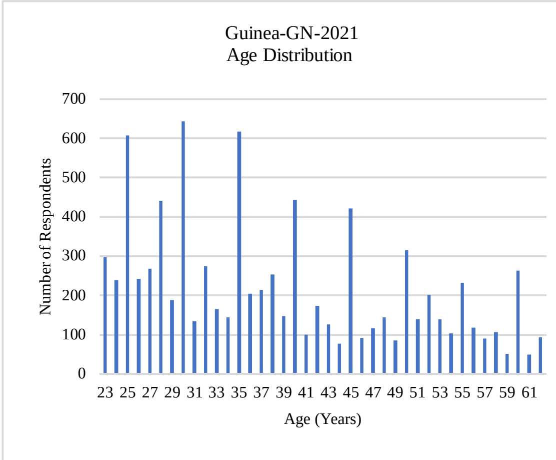
Data Source: Figures are based on Household data files of Standard DHS (ICF, 1985-2023)

Total of 59 survey data sets; 35 from Sub-Saharan Africa, 6 from North Africa/West Asia/Europe, 2 from Central Asia, 9 from South & Southeast Asia, 1 from Oceania, and 6 from Latin America & Caribbean were considered to check the quality of data. All participating countries from "central Asia" and "Latin America and the Caribbean" have perfectly accurate (Tajikistan, Guatemala, Haiti) or fairly accurate (Kyrgyz Republic, Colombia, Dominican Republic, Honduras, Peru) age data. Most of the Sub-Saharan African and South-southeast Asian countries have rough or very rough reported age data. Table 2 showed the details of all participating counties and the quality of reported age distribution.

Figure 1b. Graphical overview of age heaping In DHS surveys for selective African countries







Data Source: Figures are based on Household data files of Standard DHS (ICF, 1985-2023)

Table 2. Whipple index in most recent DHS surveys for all participating DHS countries.

Country	Data Set	Number of Individuals	Study Sample Age Range (23-62 years)	Whipple Index	Deviation from original	Data Quality
Afghanistan	**AF-DHS-2015**	**203708**	**65547**	**191**	**91.16%**	**Very poor/rough**
Albania	AL-DHS-2018	54019	28284	105	5.24%	Fairly accurate
Angola	AO-DHS-2015-16	74902	23183	128	28.10%	Poor/rough
Armenia	AR-DHS-2016	27768	15654	116	15.88%	Moderate
Bangladesh	**BD-DHS-2018**	**89819**	**40456**	**187**	**86.93%**	**Very rough**
Benin	BJ-DHS-2017-18	74673	25352	150	50.48%	poor/rough
Burkina Faso	BF-DHS-2010	820095	66989	128	27.78%	Poor/rough
Burundi	BU-DHS-2016-17	78367	26497	133	32.69%	Poor/rough
Cameroon	CM-DHS-2018	60699	21516	138	38.11%	Poor/rough
Chad	TD-DHS-2014-15	99620	29165	226	126.20%	Very Poor/rough
Columbia	CO-DHS-2015	162459	78835	110	10.03%	Moderate
Comoros	KM-DHS-2012	24499	9168	165	65.41%	poor/rough
Congo	CG-DHS-2011-12	51449	19092	108	8.32%	fairly Accurate
Congo Democratic Republic	CD-DHS-2013	95949	30936	112	12.02%	Moderate
Cote d'Ivoire	CI-DHS, 2011-12	51187	19193	128	27.83%	Poor/rough
Dominican Republic	Dr-DHS-2013	41267	18668	110	9.97%	fairly Accurate
Egypt	EG-DHS-2014	120276	55382	127	26.56%	Poor/rough
Ethiopia	**ET-DHS-2016**	**75224**	**26561**	**186**	**85.57%**	**Very poor/rough**
Gabon	GA-DHs-2012	41675	15083	106	5.81%	fairly Accurate
Gambia	**GM-DHS-2019**	**55640**	**19248**	**155**	**55.34%**	**poor/rough**
Ghana	GH-DHS-2014	43945	16939	132	32.06%	Poor/rough
Guatemala	GU-DHS-2015	102510	39993	103	2.95%	Perfectly Accurate

Guinea	**GN-DHS-2018**	**49543**	**16544**	**182**	**82.03%**	**Very Poor/rough**
Haiti	HT-DHS-2017	59547	13191	103	2.82%	Perfectly Accurate
Honduras	HN-DHS-2012		39439	108	7.60%	fairly Accurate
India	**IA-DHS-2019-21**	**2843917**	**1437924**	**150**	**50.38%**	**poor/rough**
Indonesia	ID-DHS-2017	197723	101163	101	1.01%	Perfectly Accurate
Jordan	JO-DHS-2018	93347	40926	101	1.26%	Perfectly Accurate
Kenya	KE-DHS-2022	156571	57052	131	31.09%	Poor/rough
Kyrgyz Republic	KY-DHS-2012	35805	16349	106	5.97%	fairly Accurate
Lesotho	LS-DHS-2014	40197	15954	105	5.27%	fairly Accurate
Liberia	LB-DHS-2019-20	41999	15136	126	26.06%	Poor/rough
Madagascar	MD-DHS-2021	90322	32446	135	35.14%	Poor/rough
Malawi	MW-DHS-2015	120492	38873	125	24.90%	Moderate
Maldives	MV-DHS-2016	32656	16035	119	19.11%	Moderate
Mali	ML-DHS-2018	54571	17880	149	49.19%	Poor/rough
Mauritania	MR-DHS-2019-21	73796	24257	135	34.64%	Poor/rough
Mozambique	MZ-MIS-2018	29021	9868	114	14.51%	Moderate
Myanmar	MM-DHS-2016	55584	27084	120	20.46%	Moderate
Namibia	NM-DHS-2013	41646	16751	106	6.11%	fairly Accurate
Nepal	NP-DHS-2022	57278	25896	126	25.64%	Poor/rough
Niger	NI-DHS-2012	64011	19730	204	103.80%	Very poor/rough
Nigeria	**NG-DHS-2018**	**188010**	**67742**	**191**	**90.95%**	**Very poor/rough**
Pakistan	**PK-DHS-2017-18**	**100868**	**39982**	**153**	**52.58%**	**poor/rough**
Papua New Guinea	PG-DHS-2018	83789	33983	130	30.01%	Poor/rough
Peru	PE-DHS-2012	103211	47190	105	5.12%	fairly Accurate
Philippines	PH-DHS-2022	129724	60489	100	0.54%%	Perfectly Accurate
Rwanda	RW-DHS-2020	55920	21433	104	3.96%	Perfectly Accurate
Senegal	SN-DHS-2019	41050	13643	141	41.03%	Poor/rough
Sierra Leone	SL-DHS-2019	72248	26019	161	60.65%	Poor/rough
South Africa	ZA-DHS-2016	38850	17515	99	1.37%	Perfectly Accurate
Tajikistan	TJ-DHS-2017	44916	20217	102	2.31%	Perfectly Accurate
Tanzania	TZ-DHS-2015-16	64880	22362	124	23.65%	Moderate
Togo	TG-DHS-2013	46577	16674	155	55%	Poor/rough
Turkey	TR-DHS-2018	39914	20308	119	18.84%	Moderate
Uganda	UG-DHS-2016	91167	29252	143	42.62%	Poor/rough
Yemen	YE-DHS-2013	120923	42679	194	93.88%	Very poor/rough
Zambia	ZM-DHS-2018	65454	21512	114	13.91	Moderate
Zimbabwe	ZW-DHS-2015	43706	16330	109	8.79%	fairly Accurate

Data Source: authors calculations based on Household data files of Standard DHS (ICF, 1985-2023)

For age adjustment/correction, four countries are selected with poor/rough or very poor/rough quality of reported age distribution. Following the steps described in methods age is corrected for Afghanistan, Bangladesh, Pakistan, and Nigeria. After calculating the probabilities for each terminal digit (0, 1, 2, …, 9) for the reported ages arbitrary counts are assigned to each terminal digit based on their actual reported probability such that the sum of weights for all 10 digits (0, 1, 2, …, 9) is 10,000. A 10*10 matrix is conducted assuming the assumption of rectangular distribution that each terminal digit has equal probability. All avoided digits are ranked from most avoided to least avoided digits. Preferred digit(s) were shifted to the nearest neighboring avoided digit(s). The process will be continued till each terminal digit attained the sum of 1000. For shifting the digit from the preferred digit (s) to avoided digit (s), two methods are used; Most Avoided (MA) and Least Avoided (LA). In the MA method, first of all most avoided digit gets digit(s) from the preferred neighboring digit(s) to attain column count 1000 followed by the next most avoided digit the process will continue till the least avoided digit take complete value. The matrix of 10*10 has all column counts 1000 and the row counts an arbitrary number of individuals based on the original probabilities of terminal digits of reported ages. In the LA method, at the first step least avoided digit gets digit(s) from the preferred neighboring digit(s) followed by the next avoided digit, and so on. After assigning digits at adjusted/corrected/true places the digit weights (probabilities) are calculated for each terminal digit (0, 1, 2, …,9). Using these digits weight new distribution of adjusted/corrected/true ages of the individuals is constructed. (A detailed description of the calculation of digit shift for one data set is described in Appendix-I).

Tables 3a and 3b present the corrected age distribution for four selected countries. Countries are selected based on rough and very rough quality age data. All adjustment is based on the probabilities of terminal digits of reported ages. Whipple indices are calculated for all adjusted/corrected age distributions of individuals (table 4). Results revealed that corrected age data is perfectly accurate according to UN criteria. Graphical representation of adjusted/corrected age distribution also showed no heaping at ages ending at (0 or 5).

Table 3a. Observed/reported and adjusted/corrected/true age distributions for Pakistan and Bangladesh

Current age	Pakistan (n = 39982)			Bangladesh (n= 40456)		
	Reported	Adjusted (MA)	Adjusted (LA)	Reported	Adjusted (MA)	Adjusted (LA)
20	-	-	-	-	-	-
21	-	-	630	-	-	134
22	-	90	-	-	112	-
23	1617	1654	1654	1243	1468	1468
24	1652	1774	1774	1333	1539	1539
25	2301	1394	1394	1841	1013	1013
26	1607	1669	1669	1435	1503	1503
27	1466	1675	1710	1259	1454	1454
28	1894	1688	1688	1488	1510	1513
29	1089	1597	1621	1038	1442	1442
30	2139	1527	1527	2281	1188	1188
31	888	1585	1402	890	1451	1474
32	1414	1487	1512	1357	1635	1617
33	979	1009	1009	924	1225	1225
34	917	1016	1016	889	1165	1165
35	1877	1138	1138	2464	1355	1355
36	1062	1113	1113	989	1080	1080
37	856	1027	1024	827	1088	1088
38	1262	1124	1124	1113	1143	1136
39	760	1156	1150	701	1082	1082
40	1564	1117	1117	2150	1120	1120
41	523	1027	833	561	1090	1072
42	886	937	958	847	1076	1092
43	605	626	626	560	777	777
44	496	566	566	592	791	791
45	1312	796	796	1780	979	979
46	603	638	638	817	883	883
47	641	760	742	788	977	977
48	742	661	661	816	837	832
49	394	662	567	552	796	796
50	681	486	486	1380	719	719
51	497	725	801	352	692	689
52	952	995	984	543	698	700
53	690	708	708	426	582	582
54	585	644	644	404	547	547
55	1111	673	673	1275	701	701
56	550	580	580	535	582	582
57	437	538	519	386	521	521
58	591	527	527	540	555	561
59	356	580	657	377	723	723
60	1216	868	868	1954	1018	1018
61	282	657	282	333	814	679
62	488	488	544	416	525	639
63	-	-	-	-	-	-
64	-	-	-	-	-	-
65	-	-	-	-	-	-

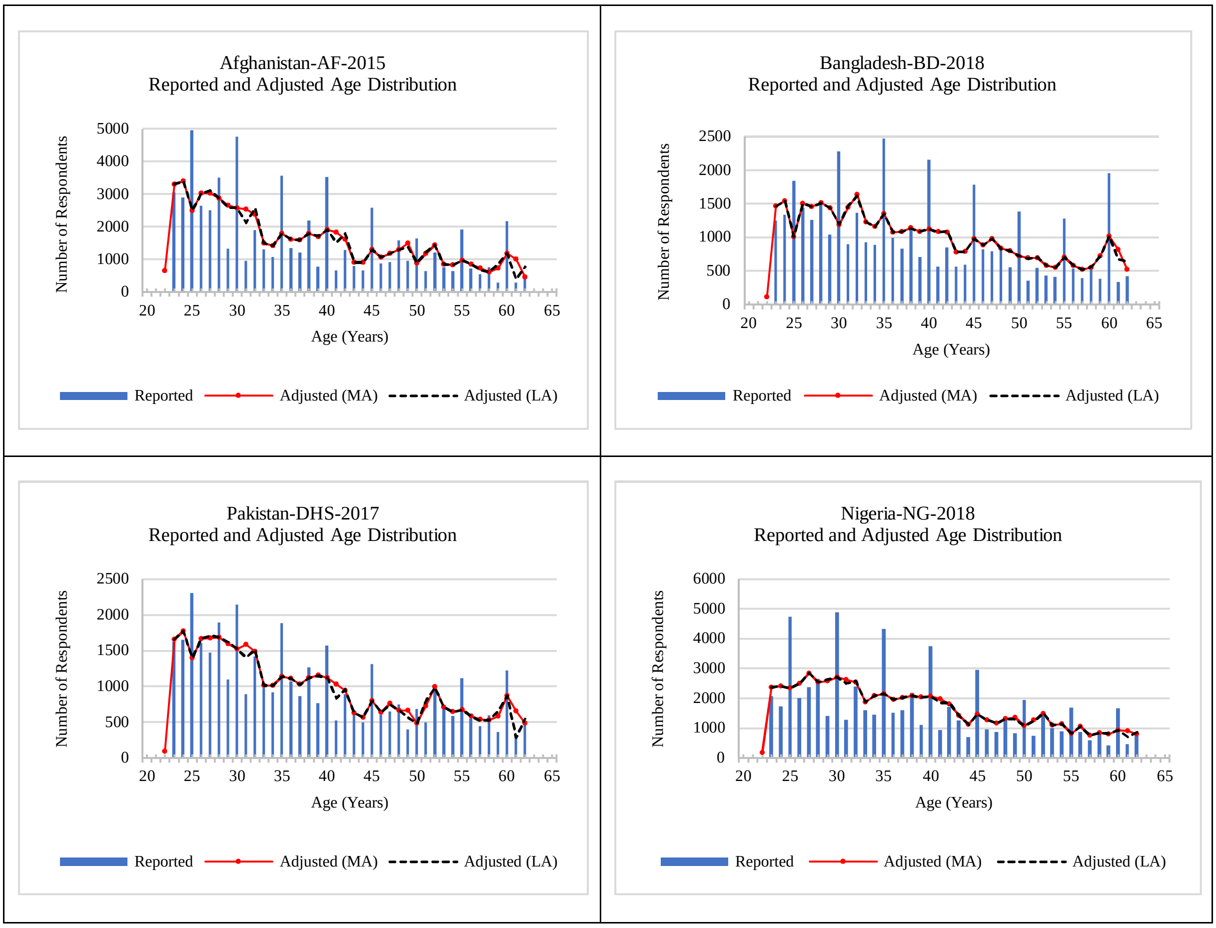
Table 3b. Observed/reported and adjusted/Corrected/true age distribution for Afghanistan and Nigeria

Current age	Afghanistan (n= 65547)			Nigeria (67742)		
	Reported	Adjusted (MA)	Adjusted (LA)	Reported	Adjusted (MA)	Adjusted (LA)
20	-	-	-	-	-	-
21	-	-	1312	-	-	507
22	-	658	-	-	166	-
23	3048	3301	3301	2054	2361	2361
24	2889	3394	3394	1713	2419	2419
25	4951	2495	2495	4736	2349	2349
26	2637	3023	3023	1995	2502	2502
27	2488	3028	3111	2372	2841	2848
28	3501	2864	2864	2643	2527	2527
29	1316	2650	2591	1404	2573	2616
30	4740	2574	2574	4886	2712	2712
31	947	2530	2118	1268	2621	2502
32	1895	2368	2573	2395	2546	2585
33	1309	1491	1490	1594	1875	1875
34	1060	1423	1423	1452	2095	2095
35	3561	1795	1795	4318	2142	2142
36	1338	1616	1616	1502	1964	1964
37	1205	1593	1594	1590	2017	2014
38	2186	1788	1788	2181	2085	2085
39	773	1685	1717	1106	2041	2032
40	3516	1910	1909	3734	2073	2073
41	658	1832	1510	936	1970	1841
42	1280	1623	1783	1697	1800	1842
43	784	915	915	1238	1429	1429
44	647	910	910	685	1124	1124
45	2578	1300	1300	2943	1460	1460
46	865	1066	1066	949	1264	1264
47	905	1186	1186	872	1163	1156
48	1579	1292	1292	1384	1323	1323
49	945	1493	1388	822	1351	1300
50	1643	892	892	1927	1069	1069
51	636	1185	1220	743	1277	1227
52	1194	1447	1429	1421	1479	1496
53	752	849	849	992	1100	1100
54	637	831	831	884	1133	1133
55	1905	960	960	1669	828	828
56	704	853	853	871	1050	1050
57	534	742	666	588	753	753
58	744	609	609	879	840	840
59	285	730	864	407	807	818
60	2165	1176	1176	1658	920	920
61	288	1011	392	444	903	706
62	459	459	768	790	790	855
63						
64						
65						

Table 4. Whipple index for adjusted/Corrected age distribution for selected countries.

Country	Whipple Index (Observed age)	Data Set	Whipple Index (Adjusted age)	Deviation from original	Data Quality
Afghanistan	191	Adjusted (MA)	101	0.96%	Very Good
		Adjusted (LA)	102	1.97%	Very Good
Bangladesh	187	Adjusted (MA)	100	0.30%	Very Good
		Adjusted (LA)	100	0.36%	Very Good
Pakistan	153	Adjusted (MA)	100	0.26%	Very Good
		Adjusted (LA)	102	1.76%	Very Good
Nigeria	191	Adjusted (MA)	100	0.28%	Very Good
		Adjusted (LA)	101	0.79%	Very Good

Figure 2. Graphical view of corrected/adjusted/true ages for some selected countries.



Discussion

Age misreporting and age heaping are critical problems in developing countries that affect all demographic estimates. Over time; due to education and awareness, ages are perfectly and fairly reported around the globe (Hussey & Elo, 1997), however, a huge poor reporting is still present in developing countries (Fayehun et al., 2020; Singh et al., 2022; Szołtysek et al., 2018). Data sets analyzed in this study also showed a poor or very rough reporting of age distribution at the time of interview in most of the developing countries. A visible age heaping can be observed on digits at a multiple of 5 or 10 (figures 1a and 1b). This type of wrong reporting of ages causes biased estimates for true population distributions. Ages can be smoothed or grouped to reduce the effect of this type of digital preference as well as random error in the data sets, but these are poor solutions if the main focus is to minimize distortions due to systematic over or under-reporting of age (Bhat, 1990). The proposed method has the potential to smooth age data as well as reduced heaping at some preferred digits like multiple of 5 or 10.

Smoothing methods have the potential to reduce age heaping but in the moving average method, we lose data at the starting and ending years of reported ages, thus the sample sizes are reduced. Other smoothing methods like; Carrier and Farrag, Karup–King–Newton, Arriaga's, and UN methods are used in the age grouping of 5 or 10 years. Grouping of age data in the interval of 5 or 10 may prevent age data from age heaping, however in many demographic studies, especially in fertility analysis grouping of ages is not a good solution. In fertility studies at each age, females have a different potential for childbearing. Therefore, proposed method has a benefit on smoothing methods as it reduced heaping at individual ages.

Shifting of digits from preferred to avoided digits can be used by different algorithms; least avoided, most avoided, random avoided, least preferred, or most preferred. Here we use the least avoided and most avoided methods to shift the preferred digit to the nearest neighboring digit. Both digit shift strategies give reliable results.

Conclusions

In this paper, we have made a simple algorithm to correct the age distribution of respondents' ages stated at the time of the interview in the presence of age heaping/ age misreporting. The proposed method to correct age heaping or AMR is based on the probabilities of the terminal digit of reported ages. Furthermore, it is significant that the value of the Whipple index of True/adjusted/corrected age distribution is within the range described by UN criteria. The beauty of the method is that it is not restricted to survey data, it is fully applicable in census data sets where age heaping creates a hurdle to reach the true distribution of any population. Secondly, the age range can vary following the rectangular distribution assumption for the terminal digit of age. Given these encouraging results, it is hoped that the method will be useful in more census and surveys with reduced errors from all over the developing world. The limitation of the method is only that it is not useful when there is a significant change in the population as a result of sudden events such as the baby boom, a lot of migration in a year(s) from some specified age group, or natural/unnatural disasters which affect population distribution.

To sum up, let's restate the merits of the proposed methods shortly. First, the method is based on the rectangular distribution assumption, in general, every terminal digit (0, 1, 2, ..., 9) has an equal probability. Second, the proposed method can be applied to any survey and census. Thirdly, the method is flexible and does not require a priori knowledge of the nature of reported age biases and the avoided ages are simply replaced with neighboring preferred ages for adjustment of the age distribution. Finally, the method is computationally simple and reliable as weights used for age adjustment are based on mathematical formulation.

References

A'Hearn, B., Baten, J., & Crayen, D. (2009). Quantifying quantitative literacy: Age heaping and the history of human capital. *The Journal of Economic History*, *69*(3), 783-808.

Arriaga, E. E. (1968). *New Life Tables for Latin American Populations in the Nineteenth and Twentieth Centuries*. Institute of International Studies, University of California. https://books.google.de/books?id=FnswAAAAMAAJ

Bachi, R. (1951). The tendency to round off age returns: measurement and correction. *Bulletin of the International Statistical Institute*, *33*(4), 195-222.

Bhat, P. M. (1990). Estimating transition probabilities of age misstatement. *Demography*, *27*, 149-163.

Caldwell, J. C. (1966). Study of age misstatement among young children in Ghana. *Demography*, *3*(2), 477-490.

Carrier, N. H. (1959). A note on the measurement of digital preference in age recordings. *Journal of the Institute of Actuaries*, *85*(1), 71-85.

Carrier, N. H., & Farrag, A. M. (1959). The Reduction of Errors in Census Populations for Statistically Underdeveloped Countries. *Population Studies*, *12*(3), 240-285. https://doi.org/10.2307/2171973

Demeny, P. G., & Shorter, F. C. (1968). *Estimating Turkish Mortality, Fertility, and Age Structure* (Vol. 1306). İstanbul Üniversitesi, İktisat Fakültesi.

Fayehun, O., Ajayi, A. I., Onuegbu, C., & Egerson, D. (2020). Age heaping among adults in Nigeria: evidence from the Nigeria Demographic and Health Surveys 2003–2013. *Journal of Biosocial Science*, *52*(1), 132-139.

Gupta, P. D. (1975). A general method of correction for age misreporting in census populations. *Demography*, *12*(2), 303-312.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hussey, J. M., & Elo, I. T. (1997). Cause-specific mortality among older African-Americans: Correlates and consequences of age misreporting. *Social Biology*, *44*(3-4), 227-246.

ICF. (1985-2023). *Demographic and Health Surveys (various) [Datasets]* Funded by USAID.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression: a self-learning text* (3rd ed.). Springer. https://doi.org/https://doi.org/10.1007/978-1-4419-1742-3

Krafft, C., Kula, E., & Sieverding, M. (2021). An investigation of Jordan's fertility stall and resumed decline. *Demographic Research*, *45*, 605-652.

Machiyama, K. (2010). *A Re-examination of Recent Fertility Declines in Sub-Saharan Africa* (DHS Working Papers No. 68. Calverton, Issue.

Myers, R. J. (1940). Errors and bias in the reporting of ages in census data. *Transactions of the Actuarial Society of America*, *41*(2), 395-415.

Myers, R. J. (1954). Accuracy of age reporting in the 1950 United States census. *Journal of the American Statistical Association*, *49*(268), 826-831.

Nasir, J. A. (2013). *Fertility transition in Pakistan: neglected dimensions and policy implications* [PhD Thesis, University of Southampton]. https://eprints.soton.ac.uk/368188/

Nasir, J. A., & Hinde, A. (2014). An Extension of Modified Whipple Index-Further Modified Whipple Index. *Pakistan Journal of Statistics*, *30*(2).

Noumbissi, A. (1992). L'indice de Whipple modifié: une application aux données du Cameroun, de la Suède et de la Belgique. *Population (french edition)*, 1038-1041.

Ntozi, J. P. M. (1978). The Demeny-Shorter and three-census methods for correcting age data. *Demography*, *15*(4), 509-521.

Pardeshi, G. S. (2010). Age heaping and accuracy of age data collected during a community survey in the Yavatmal district, Maharashtra. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, *35*(3), 391.

Poston, D. L., & Micklin, M. (2005). *Handbook of population*. Springer.

Poston, D. L., Walther, C. S., Chu, I. H. J., Ginn, J. M., Li, G. J.-K., Vo, C. H., Wang, P., Wu, J. J., & Yuan, M. M. (2003). The age and sex composition of the Republic of Korea and its provinces, 1970 and 1995. *Genus*, 113-139.

Poston Jr, D., Chu, I., Ginn, J., Li, G. J.-K., Vo, C. H., Walther, C., Wang, P., Wu, J., & Yuan, M. (2000). The quality of the age and sex data of the Republic of Korea and its provinces, 1970 and 1995. *The Journal of Gerontology*, *4*, 85-126.

Pullum, T. W. (2005). A statistical reformulation of demographic methods to assess the quality of age and date reporting, with application to the Demographic and Health Surveys. Annual Meeting of the Population Association of America, Philadelphia, March,

Pullum, T. W. (2006). An assessment of age and date reporting in the DHS surveys, 1985–2003 (DHS Methodological Reports No. 5). *J Calverton, MD: Macro International*.

Pullum, T. W. (2008). *An Assessment of the Quality of Data on Health and Nutrition in the DHS Surveys, 1993-2003*.

Ramachandran, K. (1965). An index to measure digit preference error in age data. World Population Conference,

Randall, S., & Coast, E. (2016). The quality of demographic data on older Africans. *Demographic Research*, *34*, 143-174.

Samuel, O. O. (2018). Scoring the Census Priority Table for Age Heaping and Shifting: A Study of 2006 Nigeria Population Census Result. *Annals of Reviews and Research*, *1*(2), 30-38.

Siegel, J. S., & Swanson, D. A. (2004). *The methods and materials of demography* (Second ed.). Elsevier Academic Press California.

Singh, M. (2017). Understanding digit preferences analysis of 64. *International Journal of Current Research 9*(01), 45144-45152.

Singh, M., Kashyap, G. C., & Bango, M. (2022). Age heaping among individuals in selected South Asian countries: evidence from Demographic and Health Surveys. *Journal of Biosocial Science*, *54*(4), 725-734.

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

Szołtysek, M., Poniat, R., & Gruber, S. (2018). Age heaping patterns in Mosaic data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 51*(1), 13-38.

UnitedNations. (1952). *"Accuracy tests for census age distributions tabulated in five-year and ten-year groups," in PopuIation Bulletin, No. 2* (ST/SOA/Ser.N/2).

UnitedNations. (1955). *Methods of appraisal of quality of basic data for population estimates, Manual II, Series A, Population Studies No 23* (23). U. Nations.

Yusuf, F., Swanson, D. A., & Martins, J. M. (2014). *Methods of demographic analysis*. Springer.