Draft of March 1, 2005

A Statistical Reformulation of Demographic Methods to Assess the Quality of Age and Date Reporting, with Application to the Demographic and Health Surveys

Thomas W. Pullum
Department of Sociology
The University of Texas at Austin
tom.pullum@mail.utexas.edu

Abstract:

Demographers have developed several procedures to identify systematic patterns of misreporting of ages, dates, and intervals in survey and census data. Most of these methods concern digit preference, transfers across specified boundaries, omission, and non-response. Almost all methods to identify and describe such patterns use aggregated data, such as the age and sex distribution or the distribution of time since last birth. This paper adapts, integrates, and extends these procedures within the framework of statistical methods for individual-level data, primarily logit and multinomial logit regression. This approach leads to the calculation of interpretable parameter estimates with standard errors and confidence intervals, corrected for the sample design, and the inclusion of covariates that may be related to misreporting. This research was motivated by the author's assessment of age and date reporting in all surveys conducted by the Demographic and Health Surveys project, and the paper includes illustrations from that assessment.

Acknowledgements:

The work described here was supported in part by a contract from Macro International related to the Demographic and Health Surveys project, which is funded by the United States Agency for International Development. I am particularly grateful to Fred Arnold and Martin Vaessen of DHS and Macro International. Some assistance in data preparation was provided by Khatuna Doliashvili.

Prepared for 2005 Annual Meetings of the Population Association of America, Philadelphia

A Statistical Reformulation of Demographic Methods to Assess the Quality of Age and Date Reporting, with Application to the Demographic and Health Surveys

1. Introduction

During the past several decades there has been a general shift of methodological perspective in demographic analysis. This shift has coincided with the increasing availability of individual-level data, the development of statistical methods to analyze such data, and vastly increased computing power. If one examines the principal references on demographic methods from more than thirty years ago, such as Barclay's *Techniques of Population Analysis* (1958) or Shryock and Siegel's *Methods and Materials of Demography* (1971), one finds scarcely any applications to survey data, or any differences between how census data and survey data would be analyzed. Rates, means, and proportions were calculated without consideration of standard errors or confidence intervals. Irregular distributions or missing information were handled with complex interpolation procedures and stable population assumptions. Two-way or three-way tables were used for the calculation of differences between groups rather than multivariate statistical models. If one compares articles in *Demography* of that time with those of today, one finds some mathematical modeling, and some chi-square statistics, but relatively little statistical modeling. The methods of logit regression, Poisson regression, hazard modeling, and other tools that are widely used now were simply not available then.

One category of demographic methods that has *not* participated in the transition to a statistical framework is the analysis of the quality of age and date reporting. For example, Frank Hobbs' chapter on age and sex composition in Siegel and Swanson's reissue of *Methods and Materials of Demography* (2004) includes about 20 pages (pp. 136-155) on the analysis of deficiencies in age data. The discussion is complete and up-to-date, but it shows that the techniques to identify and adjust for such deficiencies are still expressed in terms of aggregated data—principally age ratios, sex ratios, and summary measures such as Myers' Blended Index.

In the course of preparing an assessment of the quality of age and date reporting in virtually all of the Demographic and Health Surveys (DHS) surveys of households and women conducted between 1985 and 2003, I have attempted to re-state the traditional approaches in a statistical framework. The goal has been to maintain the reasoning behind these methods but to modify and extend them to have the following features:

--to be calculated from files of individual cases
--to be calculated with a readily available statistical package, such as STATA
--to be calculated with statistical models such as logit regression
--to be accompanied with parameter estimates, standard errors, confidence intervals, and test statistics
--to incorporate the sampling design by using sampling weights and clustering
--to be able to to include categorical or interval-level covariates.

This paper will present the strategy for doing this, illustrated with several measures and with results from DHS surveys.

2. General strategy

The reports of age and date in DHS surveys that are of primary interest include the following:

--ages of all household members, reported in the household survey
--ages and birth dates of women, reported in the survey of women 15-49
--ages at marriage and dates of marriage of these women
--dates of birth of children
--ages at death of children who have died, especially before age two.

DHS surveys include other dates and ages, but the ones listed above would generally be considered most important, and the others would be analyzed similarly.

Three principal kinds of problems can arise with these reports. First, in the surveys of women, although not in the household surveys, there can be incompleteness in the reporting of an age or birth date. For example, a woman may report her current age but not her birth year or birth month. In this case, or if the responses that are given are internally inconsistent, DHS uses automated imputation procedures so that age, birth year, and birth month are all present and consistent.

The second type of problem is heaping. This usually takes the form of excessive numbers of cases reported at ages ending in 0 or 5, but the heaping can occur at other digits, especially for persons under age 20, and sometimes takes the form of heaping on calendar years that end in 0 or 5. Age at the death of a child is given in months if the child died before the second birthday, with substantial heaping at 12 months and often some heaping at 6 and 18 months. This pattern is also characteristic of duration of breastfeeding and any other durations given in months, but our examples will only relate to child deaths.

Thirdly, there can be net transfers or displacements of age. (Heaping, by contrast, is the result of age transfers, but with approximately equal numbers of upward and downward transfers.) Eligible respondents for the surveys of women are identified by their age in the household roster. Interviewers have some motivation to shift the ages of women who are just inside the boundaries of the 15-49 interval in order to reduce their workload of later interviews of women.

Note that the omission of eligible cases is not included in this list of potential problems. Omission could be extremely serious, for example if there was systematic omission of births, especially of ones that had resulted in an early child death. There is very little evidence of such omission in DHS surveys, at least, so it will not be conisdered to be an issue. In some other applications, such as an evaluation of vital statistics reporting in a developing country, it certainly could not be ignored.

Incompleteness of age and date reporting is somewhat different from heaping and displacement because it is readily identified for individual cases and described by a distribution of the different kinds of information that are given (e.g. age, year of birth, and month of birth) and the degree of compatibility of these components. Imputation procedures to force compatibility or to estimate missing components, with a random component, have been developed by DHS and will not be described or assessed here.

Heaping and displacement have the fundamental difference that it is not possible to say whether a specific case has been misreported. For example, some cases with reported age 50 may have been affected by systematic heaping or displacement, but it is impossible to distinguish them with certainty from cases for which the true age is, indeed, 50.

Existing methods to detect heaping and displacement typically proceed through two steps: the calculation of expected values, followed by a summary of deviations from these expected values.

The first step involves the calculation of expected frequencies, proportions, or ratios. For example, Myers' Blended Index assumes that, after adjustment, each final digit 0 through 9 will be equally likely. The expected frequency of each digit is $n/10$, where n is the sample size. As another example, heaping at 12 months may be identified by the calculation of an expected frequency at 12 months. Sullivan, Bicego, and Rutstein (in Arnold, et al., 1990) and Marckwardt and Rutstein (1996) calculated the expected value at 12 months as the average for months 10, 11, 13, and 14. Singh (in Goldman, Rutstein, and Singh, 1985) and Curtis (1995) calculate the expected value at 12 months as the average for months 10, 11, 12, 13, and 14.

In an essentially similar way, sex ratios and age ratios to detect displacement involve the calculation of expected values. From the household schedule, the ratio of males to females can be calculated at age 10-14 and at age 15-19. If there are no downward transfers of women across age 15, one can think of either of these sex ratios as the (estimated) expected value of the other. If there are net downward transfers, they will tend to make the sex ratio at 15-19 larger than the sex ratio at 10-14. Similarly, looking just at the age distribution of females, age ratios can detect displacement. If there are no downward transfers of women across age 15, then the ratio of females age 10-14 to females age 5-9 should be about the same as the ratio of females age 15-19 to females age 10-14; either can be taken as the expected value of the other. But if females have been systematically shifted downwards across age 15, then the first ratio should be be noticeably larger than the second one.

The second step in these methods is the calculation of a summary index based on differences that should be close to zero or ratios that should be close to one if there is no misreporting. Thus, Myers' Blended Index is just the index of dissimilarity for a comparison of the observed (but blended) proportions at each final digit with the expected proportions, uniformly .10: it is one-half the sum of the absolute deviations. Rutstein and Bicego (1990) use an overall measure of age displacement which is the sum of the difference between the two sex ratios at age 15 and the two sex ratios at age 45, and a similar measure using the two age ratios at age 15 and the two age ratios at age 45. The ratio of observed to expected numbers of deaths at 12 months has been used in DHS assessments.

All of the procedures described above are amenable to a translation into standard statistical methods, although, to our knowledge, this has not yet been done. Incompleteness is easiest to translate, because it is readily expressed as a binary outcome, for which logit regression is available. Heaping and displacement can also be analyzed with logit and multinomial logit regression, although the definitions of the outcome variables may be less obvious. In a sense, many statistical methods are intended for the calculation of individual-level expected values and deviations of observed values from expected values, and it is perhaps surprising that they have not been used more to assess the quality of age and date reporting.

3. Data

As mentioned in the introduction, the impetus for this paper was an opportunity to conduct an assessment of the quallity of age and date reporting in DHS surveys of households and women. The report on that assessment (Pullum, 2005) includes 125 out of the 141 surveys conducted from 1985 to 2003. Sixteen surveys were omitted because they were restricted or because the standard recode files were not yet available. DHS-I household surveys (1985-89) were never put into standard recode format and are omitted.

This analysis is thus based on 99 household surveys conducted between 1990 and 2003 and 125 surveys of women age 15-49 conducted between 1985 and 2003. For each survey of women, a file of all their children was constructed. Sixty-one countries are represented, with one to five surveys from each country. Table 1 gives a brief description of the data files in terms of the numbers of cases. All three kinds of files will appear in our examples. Any results reported from the DHS household surveys will be limited to de jure residents (hv103=1). File preparation and data analysis were done with Stata, version 8.

Table 1. Number of data files used and number of cases in those files.

	Number of files	Smallest	Median	Mean	Largest
Household files	99	5,764	36,885	69,143	488,839
Woman files	125	1,286	7,070	9,904	90,303
Child files	125	3,256	24,357	30,865	285,599

Notes:
The household files consist of one record for each person in each household
The smallest files are all from the Dominican Republic 1999 survey
The largest household and woman files are from the India 1998/99 survey
The largest child file is from the India 1992/93
survey

4. Statistical analysis of incompleteness of ages and dates

The survey of women age 15-49 includes the following ages and dates for which the analysis requires the construction of a century-month code (cmc), which in turn requires both a calendar year and a month:

--ages and birth dates of women
--ages at marriage and dates of marriage of these women
--dates of birth of children.

For each of these events, DHS provides a variable that tells whether there was any inconsistency or incompleteness in the original data, in which case a reconciliation or imputation procedure was used to generate the age, year, month, and cmc that appears in the standard recode file. Here we shall refer to any kind of modification of the original responses as imputation. We will not describe the imputation procedure except to say that it is very sophisticated and includes a random component to avoid any additional heaping on age, year, or month. To assess the quality of the data, the most important distinction is whether or not any imputation was required.

The standard way to describe the level of incompleteness would be with a proportion or a percentage, perhaps calculated within subgroups, but the natural statistical model for a binary outcome such as this is logit regression. A variable y, "incompleteness", will be assigned the value 0 if reporting of age and birthdate was complete (v014=1) and 1 otherwise. We could obtain the proportion of incomplete responses simply with a logit regression of y with no covariates, getting a coefficient b_0 on the logit scale. That is, we calculate $\log it(\hat{y}) = b_0$. ($\log it(\hat{y})$ is a shorthand notation for $\log[\hat{\Pr}(y=1)/\hat{\Pr}(y=0)]$.) The exponential of b_0 will be the observed odds of an incomplete response, and the observed proportion will be given by $p = \exp(b_0)/[1+\exp(b_0)]$. A confidence interval for the population proportion is obtained by applying the same transformation to the two ends of the confidence interval for the population value of b_0. Then y can be regressed on covariates for a much more complete description of the pattern of incompleteness than would otherwise be possible.

This application of logit regression will be illustrated with one of the DHS surveys. The survey with the highest level of imputation of the woman's age and birth date (indicated by v014) was the 2000 survey of Bangladesh. In that survey, this information was complete for only 6.1% of the women. 81.9% of the women gave their current age, without a birth year or birth month, and another 9.6% gave their current age and a birth year, but no birth month. The remainder gave age and either month or year, but they were inconsistent. Unless all of the original information is wrong, the error induced by imputation itself will rarely exceed a year in these kinds of cases. Nevertheless, the high level of incompleteness of age and birth data in this survey is noteworthy because it probably reflects the quality of the information that *is* provided.

In a logit regression of y with no covariates we obtain $b_0 = 2.6084$, and the lower and upper ends of its 95% confidence interval are 2.4603 and 2.7564, respectively. These numbers convert to an observed proportion $p = .931$, with a 95% confidence interval of (.921, .940).

The 2000 survey of Bangladesh included 10,544 women. The file includes 171 women age 10-14, which is very unusual for a DHS survey but was part of a strategy to combat downward transfers from 15-19 to 10-14, a topic to be discussed below. Likely predictors of misreporting of all kinds in this survey are type of place of residence (urban/rural, with rural as the reference category), district (with Dhaka District as reference category), age (categorized, with 15-19 as reference category) and education (years of schooling, v133, 26 missing). Education of women in Bangladesh has a skewed distribution; 45% of the respondents had no schooling at all.

Table 2 describes a sequence of three logit regressions. Model 1 contains only type of place and district, model 2 adds age of respondent, and model 3 adds education of respondent. All four variables are highly significant in every model in which they appear. Incompleteness is most severe for women in rural parts of Dhaka District, age 35 and above, with no education. Education is clearly the most important predictor. It is very important in its own right and it explains most of the urban/rural differential. Model 3 accounts for 29% of the total deviance.

Table 2. Logit regressions of incompleteness of age and birthdate
Reporting on type of place of residence, district, age, and education.
DHS survey of Bangladesh 2000. OR: Odds Ratio; z: test statistic

Incomplete Age/Birthdate	Model 1 OR	Model 1 z	Model 2 OR	Model 2 z	Model 3 OR	Model 3 z
Type of Place						
urban	.29	-9.26	.28	-9.22	.68	-3.31
rural	1.00	----	1.00	----	1.00	----
District						
Barisal	.48	-2.96	.47	-2.97	.51	-2.87
Chittagong	.79	-1.10	.79	-1.07	.97	-0.18
Dhaka	1.00	----	1.00	----	1.00	----
Khulna	.37	-4.94	.38	-4.83	.32	-5.62
Rajashahi	.68	-2.02	.68	-2.01	.56	-2.98
Sylhet	.64	-1.59	.61	-1.73	.50	-3.02
Age						
10-14			.77	-0.86	.57	-1.80
15-19			1.00	----	1.00	----
20-24			.82	-1.49	.87	-0.99
25-29			.83	-1.35	.74	-2.04
30-34			1.25	1.36	1.21	1.16
35-39			2.21	4.61	1.62	2.61
40-44			1.71	2.91	1.06	0.29
45-49			1.94	3.25	1.12	0.54
Years of Education					.69	-16.52
Pseudo R^2	.0516		.0665		.2904	

Another logit regression using the Bangladesh 2000 survey looked for variation across interviewers (v028). This survey had 85 interviewers, but 20 of them did 10 or fewer interviews. By dropping those 20 interviewers, we drop only 45 cases. The remaining 65 interviewers did an average of 161.5 interviews each. Most interviewers worked in both urban and rural areas and in more than one district, reducing the typical confounding between effects of interviewers and effects of type of place and district. In a fixed effects model (treating v028 as a categorical variable with 65 categories) there are relatively few significant differences between interviewers. 4.54% of the total deviance is explained. This is highly significant, but less than the 5.16% explained by type of place and district in model 1. Interestingly, there is a highly significant difference between the 65 main interviewers and the 20 who collectively did only 45 interviews. The mean level of incompleteness for the group of 20 interviewers was 73%, compared with 92% for the group of 65. We speculate that the 20 were supervisors. This difference suggests that incompleteness can be reduced, probably with additional probing. Of course, there is no guarantee that a more complete report is more accurate in such a context, because most people simply do not know their birthdates.

5. Statistical analysis of age and date heaping

5.1 Heaping on a single value

Sometimes there is evidence that a specific number in a relatively short sequence appears more often than it should. For example, one way to assess upward transfers out of the 15-49 age range of eligibility for the survey of women would be to focus on the frequency of age 50 within a range of ages close to 50. To assess backward transfers in the birth history one could compare the reported number of births in the calendar year immediately before the window of health questions with the surrounding calendar years. Another example, and the one we shall focus on, is heaping of children's deaths at 12 months. The first two examples relate to transfers, with heaping interpreted as a symptom of transfers.

In the first major assessment of DHS data (Demographic and Health Surveys, 1990) and in Marckwardt and Rutstein (1996), three-point or five-point ratios (our labels, not theirs) were used to assess these three specific kinds of heaping. These ratios are defined as follows:

Three-point ratio: Let *a*, *b*, and *c* be the number of events in three consecutive intervals (e.g. ages 49, 50, 51). The "correct" value of *b* is estimated to be $(a+c)/2$; the ratio of observed to "correct" is $r = \dfrac{b}{(a+c)/2} = \dfrac{2b}{a+c}$.

Five-point ratio: Let *a*, *b*, *c*, *d*, *e* be the number of events in five consecutive intervals (e.g. ages 48, 49, 50, 51, 52). The "correct" value of *c* is estimated to be $(a+b+d+e)/4$; the ratio of observed to "correct" is $r = \dfrac{c}{(a+b+d+e)/4} = \dfrac{4c}{a+b+d+e}$.

For both the three-point and five-point ratio, the relative excess in the middle frequency is described with $r-1$ or $100(r-1)$. If this is close to 0 then the progression across the three or five intervals is close to linear. If there actually is heaping, then the average of the surrounding frequencies is not a good estimate of the correct middle frequency, but a comparison of the observed middle frequency to that average will be a good indicator of heaping, with the advantage that the observed and correct frequencies will be statistically independent.

It appears that this method has always been used with aggregated data, e.g. a tabulated distribution of single year of age or calendar year of birth or age at death in months, but it can readily be put into a logit regression format. This will be illustrated with the five-point ratio and age at death 12 months. Define an individual-level outcome *y*, age at death in months, and assign the value 1 if the reported age is 12 months, the value 0 if it is 10, 11, 13, or 14 months, and missing otherwise. We could obtain the ratio *r* simply with a logit regression of *y* with no covariates, getting a coefficient b_0 on the logit scale. The exponential of b_0 , when multiplied by four, will be the ratio *r*, that is, $r = 4\exp(b_0)$. This procedure will give a test statistic for the null hypothesis that the ratio is one (i.e. that b_0=0) in the population. A confidence interval for the population value of *r* is obtained by exponentiating the confidence interval for the population value of b_0 and multiplying by four. As in section 4, logit regression will also allow for sampling weights and clustering and will allow the inclusion of covariates.

The application to DHS data will be based on deaths to children born in the ten years before the survey. Going back further than this simply exacerbates the heaping at 12 months. The four DHS surveys with the most severe heaping of age at death at 12 months are the 1988 survey of Ghana, the 1987 survey of Mali, the 1987 survey of Burundi, and the 1999 survey of Guinea. In general, as suggested by the dates of these surveys, there has been a substantial reduction over time in this kind of heaping. This example will use Guinea 1999.

Using the child file for the Guinea 1999 survey, a logit regression of *y* with no covariates has an intercept 1.6244 and 95% confidence interval (1.2818, 1.9669) for the population value of the intercept. Exponentiating each of these numbers and then multiplying by 4, the observed five-point ratio is 20.301 and a 95% confidence interval for the population value of the five-point ratio is (14.413, 28.595) This is a wide interval because the frequencies are rather small. The survey has 17 (unweighted) deaths at 10 months, 10 at 11 months, 245 at 12 months, 9 at 13 months, and 12 at 14 months. Despite the large standard error, the five-point ratio is very significantly greater than 1.

Additional logit regressions do not show significant covariation by type of place of residence (urban, rural), sub-national region (Lower Guinea, Central Guinea, Upper Guinea, Forest Guinea, Conakry) or age of mother. However, education of the mother has a significant negative effect on the ratio. Of these 293 child deaths, 264 were to women with no schooling at all, and the remaining 29 were to women whose education varied from one year to eleven years. The coefficient for education is -.0427. If this is exponentiated (it is *not* necessary then to multiply by 4), we get an estimated reduction of 4.2% in the five-point ratio per year of education. The robust standard error of this coefficient is .0168, and a one-tailed *z* test statistic is significant at the .01 level. (The test is one-tailed because we would hypothesize that education has a negative effect on heaping.)

Some final cautions must be made. First, it is well known that child mortality tends to cluster by households and women. These 293 child deaths in this data set did not all occur to different women: 19 of the mothers had two child deaths. A correction for this would increase the standard errors of both the intercept and the slope of the logit regression. Second, the extreme skew in the distribution of the mother's education may undermine the validity of a test of the education effect. However, because of the small sample size we will not push this example any further.

Severe heaping of age at death 12 months may lead to an under-estimate of infant mortality (at age 0) and an over-estimate of child mortality (at age 1-4) because some of those mis-reported deaths actually occurred in months 0-11. DHS sometimes re-distributes one-fourth of the excess at 12 months back into the interval 0-11 months. I have also attempted to develop procedures to re-distribute heaped births but will not go into them here.

5.2 Digit heaping

In many settings there is a pronounced tendency to give ages, particularly of adults, that end in the final digits 0 or 5. This is a natural result of having a number system with base ten and not being confident of one's true age. Heaping on multiples of 6 and 12 when reporting durations in months, or heaping on multiples of 7 when reporting durations in days, etc., is common when providing an estimate.

In order to identify the amount of age heaping, the first step would be to look at the relative frequency of each final digit, 0, 1, 2, …, 9. However, most age distributions have fewer and fewer cases as age increases (particularly in terms of true age), because of the combined effects of mortality and a history of population growth. As a result, there will tend to be more cases with final digit 0 than final digit 1; more cases with final digit 1 than final digit 2, and so on. One could imagine many ways to adjust for this pattern, and several have indeed been proposed, but the most common measure of age heaping used by demographers is Myers' Blended Index. It requires that the full range of ages be a multiple of ten, such as ages 10 to 89 (which includes 89-10+1=80=8*10 years), and gives weights to each age in the following pattern. Say that the starting age is 10 and the final age is 89. Then (see, for example, Siegel and Swanson, 2004, pp. 138-9) the weight for age 10 will be 1; for age 11 will be 2; …; for age 18 will be 9; and ages 19 through 89 will have weight 10. The age range for Myers Index does not have to begin with 10; the main reason why it usually does is that less heaping is usually observed in the early ages.

We will not go into the reasoning behind these weights, but they in practice they seem to work well. One uses them to calculate the weighted sum of cases with reported final digit 0, the weighted sum for final digit 1, etc., and converts these ten sums into percentages of the weighted grand total. If there were no heaping, each percentage would be 10. The final index is calculated as half the sum of absolute deviations from 10. This is a form of an index of dissimilarity and can be interpreted as the percentage of cases that would have to be transferred from one final digit to another in order to achieve a uniform (adjusted) distribution across digits.

The measure would detect any kind of widespread digit preference, e.g. for even numbers, not just a preference for 0 and 5.

Although Myers' Index is traditionally calculated from aggregated data, that is, from an age distribution in single years of age, using a spreadsheet, it can also be calculated from individual-level data using multinomial logit regression. Within a range such as 10-89, a respondent's age is converted to a tens digit and a ones digit, which may be called *digit10* and d*igit1*, respectively. The outcome variable *y* is *digit1*, which has ten "categories", labeled 0 through 9. The weights, described above, are a variable *wt*. A multinomial logit regression with dependent variable *y*, no covariates, and frequency weights *wt* is calculated. (In Stata, the command is simply "mlogit y [fweight=wt]".) Then (using the "predict" command in Stata) one constructs ten variables which are the estimated probabilities that y=0, y=1, …, y=9 for each case in the file. The estimated probabilities will be the same for every case and may be referred to as $\hat{p}_0, \hat{p}_1, ..., \hat{p}_9$.

One then combines these ten variables into the variable $M = 50\sum_{i=0}^{9} | \hat{p}_i - .1 |$ to get Myers' Index. (The factor of 50 comes from the conversion to percentages and then division by 2.) *M* will be attached to every case in the file but will have the same value for every case. Myers Index itself could be retrieved by listing the value of *M* for the first case in the file, say, or calculating the mean of *M*.

Unfortunately, standard errors and test statistics will be distorted by the weights 1 through 10 used in the blending process, because they artificially inflate the size of the sample. This distortion will be greatly reduced by treating the Myers weights like sampling weights (pweights rather than fweights). In Stata, the pweight option always normalizes the weights to have an average of one. Since the Myers weights are ten for most cases, pweight will approximately have the effect of dividing the Myers weights by 10, resulting in a weight near one for most cases. The cases in the first nine years of the full range, especially the first four or five years, will be largely discounted in the calculation of standard errors, but they usually show the least evidence of heaping. If the survey includes sampling weights, as most DHS surveys do, then the Myers frequency weights would be multiplied by the sampling weights, and the product would be treated as pweights for the calculation of robust standard errors.

Putting aside the effect of the Myers weights, a test of the null hypothesis that *M*=0 in the population would be equivalent to a simultaneous test of the null hypothesis that all the constants or intercepts in the multinomial logit regression are zero. Calculation of the standard error of *M* in order to construct a confidence interval for the population value or to test other hypotheses would require some other procedure, such as a bootstrap.

The real payoff from the multinomial logit framework is that it allows the incorporation of interval-level or categorical covariates. For example, *w* could be a covariate such as years of education of the household respondent. The procedure is essentially the same as above, with the following changes. First, *w* would be included in the multinomial logit regression. Second, the variable M would become *M*(*w)*, a function of *w*. This function is not necessarily linear and may not even have a continuous first derivative because of the role of the absolute values. However, the function is easily graphed, and the effect of *w* on heaping can be assessed, at least in terms of

magnitude and statistical significance (with the caveats above), from the change in the log likelihood of the of the multinomial logit regression when *w* is added.

To illustrate individual-level and multivariate version of Myers' Index we will use the 1998/99 survey of India. The 1990 survey of Nigeria and the 1991 survey of Pakistan have slightly higher levels of heaping than this survey, but they are used in other examples in this paper. Evidence of age heaping is equally strong in the 1993 survey of India. India 1998/99 is the largest survey DHS has ever carried out. There were a total of 488,839 de jure residents age 0-79 in the household survey, the basis for these calculations. Males and females are combined. Table 3 gives the distribution across the final digit *y* in three forms: unweighted; weighted by the sampling weights in the survey; and weighted by the product of the Myers weights and the sampling weights. The third column can be used directly for the calculation of Myers' Index by summing the absolute deviations from 10% and dividing by two. This calculation gives *M*=17.1. There is very pronounced heaping at final digits 0 and 5, a small amount of heaping at 2 and 8, and avoidance of the other six digits. In order to achieve a uniform (adjusted) distribution, 17.1% of respondents would have to be shifted from digits 0, 2, 5, or 8 to the other six digits.

Table 3. Percentage distribution of household residents across final digit of age 0-9. DHS survey of India, 1998/99.

Column (1): Unweighted
Column (2): Weighted by sampling weights
Column (3): Weighted by product of Myers weights and sampling weights

y	(1)	(2)	(3)
0	18.28	18.28	18.10
1	7.04	7.16	6.06
2	11.11	11.17	10.73
3	7.58	7.51	6.91
4	8.03	8.01	7.57
5	16.37	16.43	17.17
6	8.59	8.63	8.78
7	6.95	6.90	7.18
8	10.27	10.21	11.12
9	5.79	5.71	6.39
Total	100.00	100.00	100.00

The multivariate form of the model will be illustrated with two covariates. We first define *w* to be the tens digit of the reported age, which takes the values 0, 1, …, 7. This covariate provides a way to quantify the commonly observed increase in heaping as age increases. Figure 1 shows this pattern very clearly. (*M*(*w*) is only defined for integer values of *w* and the graph connects the fitted values.) Heaping is not negligible in the first two decades of age but is relatively low (the single year of age distribution shows heaping at ages 5, 8, 10, 12, and 18). *M* increases by an average of about 8% in every subsequent decade of age. Figure 1 (and Figure2) also shows the overall value of Myers' Index with a horizontal line at 17.1%.

Secondly, we define *w* to be the completed years of schooling of the household respondent, the person who provides the information about the household, including the ages of household members. This person is usually the household head or the spouse of the head. 45% of the household respondents have no formal schooling at all, and only 10% have 12 or more years of schooling.

Figure 1. Myers' Index by tens digit, India 1998/99.

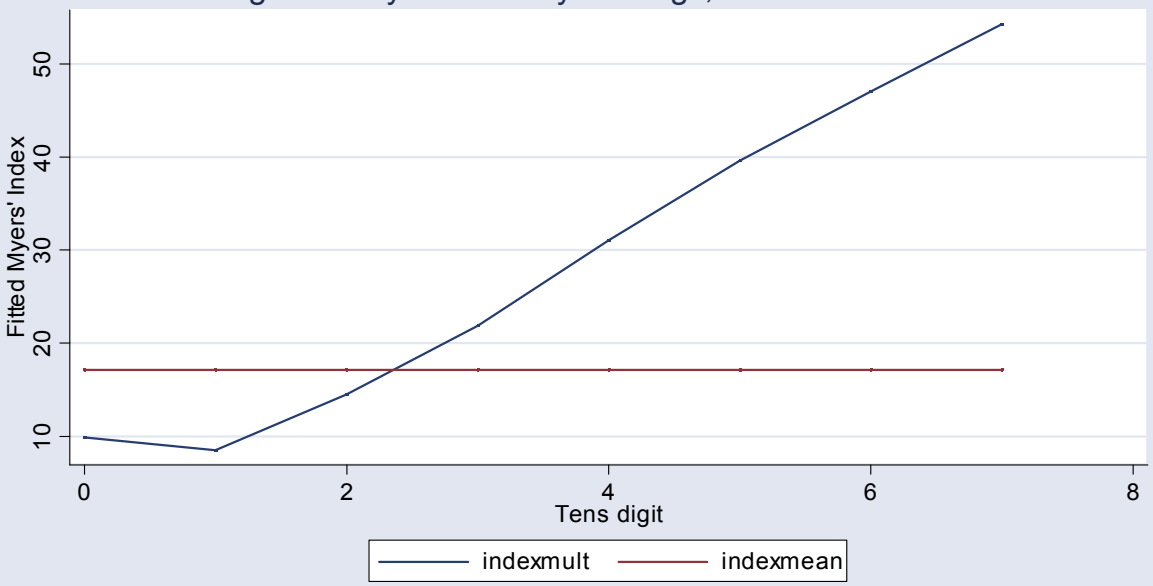
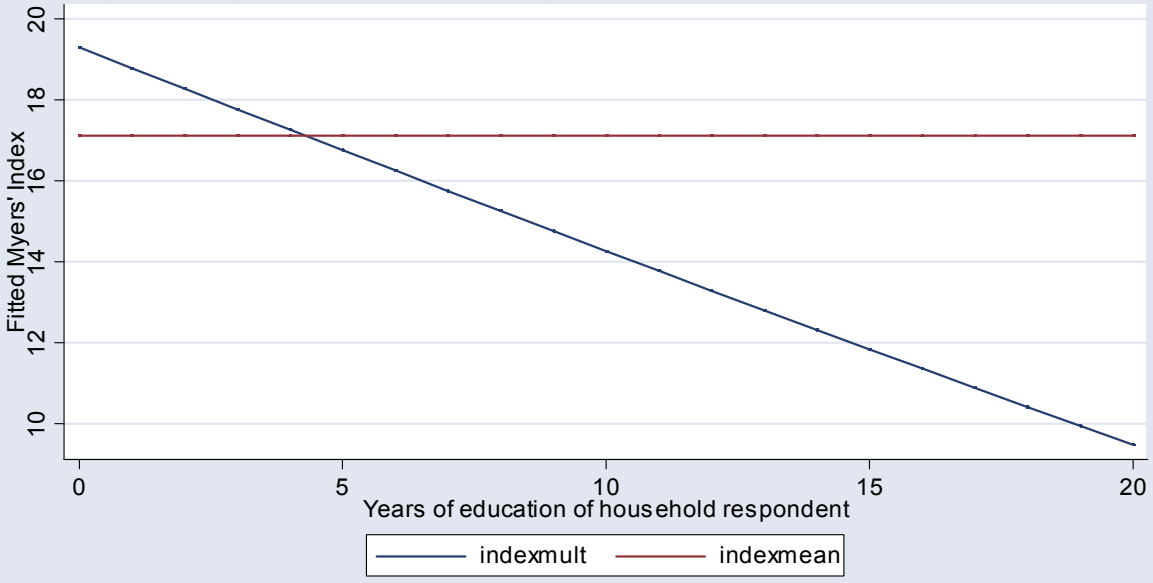


Figure 2. Myers' Index by schooling of household respondent, India 1998/99.



This variable has a highly significant negative effect on heaping. The Wald chi-square is 1439 with 9 degrees of freedom. The fitted value of $M(w)$ is about half as large at the maximum value of respondent's schooling as at the minimum value. However, the proportion of total deviance explained is miniscule; the pseudo-R^2 is only .001.

Finally, this multivariate version of Myers' Index could lead to an alternative form of the initial adjustment with the weights 1 through 10 in the first decade. Instead of using weights, the overall shape of the age distribution could be taken into account with a statistical control rather than weights. We have explored several ways to do this but will not describe them here.

6. Statistical analysis of age displacement using age ratios

Two kinds of systematic age displacement are likely in DHS surveys. The first is a tendency for ages in the household survey to be displaced across the boundaries of eligibility for surveys of individuals. Related to the survey of women, there may be some displacement of women downward across age 15 or upward across age 50. If there is a survey of men with fixed age boundaries, there may some displacement of men downward across the lower boundary or upward across the upper boundary. This is commonly believed to be the result of interviewers seeking to reduce their workload, but it is probably more likely if the household respondent is not fully confident about the ages of household members. We will assume that such transfers are within five years of the boundary, in terms of both the true age and the reported age.

The second type of transfer, also motivated by the interviewers' possible desire to reduce their workload, is the misreporting of the ages or birthdates of children in order to avoid the detailed questions about child health. These questions are to be asked about children born since January of some calendar year—that is, eligibility is based on date of birth, not age at interview. In all surveys in DHS-I, for example, the questions were asked about all children born since January of the fifth calendar year before the year in which the fieldwork began. For example, if fieldwork began in October 1988, eligibility would begin in January 1983. In such a survey, some birthdates might be shifted backwards. We will assume that such transfers are within one year of the boundary, in terms of both the true birthdate and the reported birthdate. Thus, in this example, some births that occurred in 1983 might be transferred to 1982. In addition to distorting measures of child health, such transfers have the potential to distort fertility rates, leading to exaggerated evidence of fertility decline.

In some countries, such as Pakistan (Pullum, 1990), there is a well-documented pattern of shifting children's ages upwards in any type of data collection. The methods described here are unable to detect such a pattern or to detect transvers that occur in a wider range than ± 5 years for adults and ± 1 year for children.

6.1 Sex ratios

Original formulation

One approach to the identification of transfers of women across age 15, for example, would be based on a comparison between the sex ratios (the number of males divided by the number of females, sometimes multiplied by 100) for ages 10-14 and 15-19. If there is no net displacement, we would expect these two sex ratios to be approximately equal.

Looking at the Botswana 1988 survey, Rutstein and Bicego (1990) calculated the following sex ratios:

Age Interval	Sex Ratio
10-14	85
15-19	102
.....	
45-49	107
50-54	46

These numbers suggest a net deficit of women age 15-19 and age 45-49, and an excess of women age 10-14 and age 50-54, consistent with a hypothesis of transfers out of the 15-49 interval of eligibility. Rutstein and Bicego summarize this indication of displacement with SR_o-SR_i, the sex ratio outside (o) the boundary minus the sex ratio inside (i) the boundary. For Botswana 1988, this indicator is 85-102 = -17 for the lower age boundary and 46-107 = -61 for the upper age boundary, or a total of (-17) + (-61) = -78. The more negative this total, the greater the indication of transfers in the hypothesized direction.

Revised aggregate-level approach

This approach can be re-stated as a statistical model. Focusing on the lower boundary for eligibility, age 15, we identify one age interval above the boundary and one below it, for both males and females:

	Males	Females	
Age Interval	Observed Frequency	Observed Frequency	Fitted Frequency
10-14	a	b	$\hat{b}$
15-19	d	c	$\hat{c}$

Thus, for Botswana 1988, a/b=.85 and d/c=1.02.

An assumption that the sex ratios should be the same for these adjacent age intervals is equivalent to an assumption that the age ratio for females should be the same as the age ratio for males. That is, if $a/b=d/c$, then $c/b=d/a$, or vice versa. We can then obtain estimates of b and c,

i.e. $\hat{b}$ and $\hat{c}$, by assuming that $\hat{b} + \hat{c} = b + c$ and that $\hat{c}/\hat{b} = d/a$ (or, equivalently, that $\ln(\hat{c}/\hat{b}) = \ln(d/a)$).

We now define two somewhat artificial binary variables, *y* and *x*. The outcome variable *y* is coded 1 for the five-year age group just above the boundary age (e.g. 15 or 50) and 0 for the five-year age group just below; otherwise it is missing. The variable *x* is coded 1 for females and 0 for males. The cross-tabulation of *y* and *x* is shown in table 3.

Table 3. Layout of two age groups, and sex, into a 2x2 table for logit regression approach to age tansfers based on sex ratios. *y* is 1 for the five-year age group above the boundary and 0 for the five-year age group below the boundary; *x* is 1 for females and 0 for males. *a, b, c,* and *d* are the numbers of cases in the four combinations.

		x = 0 Male	*x* = 1 Female
y = 0	(first age group)	*a*	*b*
y = 1	(second age group)	*d*	*c*

We then do a logit regression of *y* on *x*. (In Stata, if the generic label for the frequencies is *n*, the command would be logit y x [fweight=n].) The results would include two coefficients b_0 and b_1, with $\log it(\hat{y}) = b_0 + b_1 x$. For aggregated data, these would be equivalent to $b_0 = \ln(d/a)$, the age ratio for males, and $b_1 = \ln(c/b) - \ln(d/a) = \ln(c/b) - \ln(\hat{c}/\hat{b})$, the age ratio for females minus the age ratio for males. We define a measure of transfer, *t*, on the logit scale as simply $t = b_1$. If this measure is positive, there have been net upward transfers, as would be expected when the boundary age is 50. If it is negative, then there have been net downward transfers, as would be expected when the boundary age is 15. The standard error of *t* is simply the standard error of b_1, and the output for the logit regression will give a *z* test statistic for a test of the null hypothesis that *t* is 0 in the population and a 95% confidence interval for the population value of *t*. Estimates of the probabilities of downward and upward shifts can be calculated in the same way that will be described below for estimates of transfers based on age ratios within the female sub-population.

Exactly the same transfer coefficient would be obtained by regressing x on y. Then the intercept would be the log of the sex ratio in the first age group and the slope would be sex ratio in the second age group minus that in the first age group. This would be more consistent with the reasoning behind the original formulation, but since misreporting is an issue for age but not for sex, the implied causal structure would be problematic. Also, additional covariates would not make sense.

The logit regression of y on x can readily be extended to individual-level data and to include covariates. However, the sex ratio approach is only valid when there is no male survey, because if there is a male survey there will probably be age tansfers for men as well as women. DHS often has a male survey, especially in recent years, so we will not actually employ the sex ratio approach, but its simplicity serves as a good introduction to the age ratio approach, which *will* be used.

6.2 Age ratios

Original formulation

We now describe the original formulation of the "age ratio" approach. The age ratio for an age interval is defined to be the ratio of the reported number of cases in that interval, divided by the reported number of cases in the preceding age interval (sometimes multiplied by 100). Because of a past history of population growth in almost all countries, and the increase in mortality with age, we expect age ratios to be only a little less than one (or 100) at most ages and to decline fairly regularly as age increases. Looking for possible transfers outside the 15-49 interval, Rutstein and Bicego (1990, p.8) reported the following relevant age ratios from the 1988 household survey of Botswana:

Age Interval	Age Ratio
10-14	126
15-19	76
.....	
45-49	73
50-54	152

The irregularity of these age ratios clearly indicates displacement from 15-19 to 10-14 and from 45-49 to 50-54. To summarize this displacement, Rutstein and Bicego calculated AR_i-AR_o, defined as the age ratio inside (i) the boundary minus the age ratio outside (o) the boundary. For Botswana 1988, this indicator is 76-126 = -50 for the lower age boundary and 73-152 = -79 for the upper age boundary, or a total of (-50) + (-79) = -129. We will build upon the logic of this approach, first with aggregated date, and will then adapt it to a statistical model that can be used with individual-level data.

Revised aggregate-level approach

Focusing on the lower boundary for eligibility, age 15, we identify two age intervals below the boundary and two above it:

Age Interval	Observed Frequency	Fitted Frequency
5- 9	a	a
10-14	b	$\hat{b}$
15-19	c	$\hat{c}$
20-24	d	d

Thus, for Botswana 1988, b/a=1.26 and c/b=.76; frequency d was not used by Rutstein and Bicego. We propose a model to estimate the second and third frequencies (b and c) using the first and fourth frequencies (a and d), which are assumed to be reported correctly. This will be done with two assumptions. The first is that the only error of reporting is in the allocation across the second and third intervals. That is, the first and fourth frequencies are "correct", $\hat{a} = a$ and $\hat{d} = d$, and the sum of the middle two frequencies is "correct", $\hat{b} + \hat{c} = b + c$. Thus it is important that the intervals be wide enough for it to be plausible that displacement does not extend beyond the middle two intervals.

The second assumption is that the "correct" frequencies have a linear pattern of progression on a log scale (most models for frequencies or odds assume some form of linearity on a log scale). That is, it is possible to find constants α and β such that $\ln(\hat{b}/\hat{a}) = \alpha$, $\ln(\hat{c}/\hat{b}) = \alpha + \beta$, and $\ln(\hat{d}/\hat{c}) = \alpha + 2\beta$.

It is not actually necessary to solve for the two constants, because the sum of the three equations gives

$\ln(\hat{d}/\hat{c}) + \ln(\hat{c}/\hat{b}) + \ln(\hat{b}/\hat{a}) = 3\alpha + 3\beta = 3(\alpha + \beta)$, or $\ln(\hat{d}/\hat{a}) = 3\ln(\hat{c}/\hat{b})$. Therefore

$\hat{c}/\hat{b} = (\hat{d}/\hat{a})^{1/3} = (d/a)^{1/3}$, which leads to

$$\hat{c} = (\hat{b} + \hat{c})\frac{(d/a)^{1/3}}{1 + (d/a)^{1/3}} = (b + c)\frac{(d/a)^{1/3}}{1 + (d/a)^{1/3}} \text{ and } \hat{b} = b + c - \hat{c}.$$

A coefficient for the amount of transfer could be calculated as the log of the observed ratio in the middle two categories minus the log of the expected ratio:

$$t = \ln(c/b) - [\ln(d/a)]/3 = \ln(c/b) - \ln(\hat{c}/\hat{b}) - \ln(\frac{c/b}{\hat{c}/\hat{b}}).$$

With the usual Poisson-based approximations, t has an estimated variance

$s^2 = (1/9a) + (1/b) + (1/c) + 1/9d)$.

A simple test of the null hypothesis that there are no net transfers between the middle two categories would be given by $z = t/s$, which would have an approximately normal sampling

distribution if the null hypothesis is true. (However, see the cautions given below regarding the interpretation of a test statistic.)

If $b < \hat{b}$ then the proportion of cases in category 2 that are estimated to be shifted out of that category and into category 3 will be

$$p_b = (\hat{b} - b)/\hat{b} = 1 - (b/\hat{b}).$$

Similarly, if $c < \hat{c}$ then the proportion of cases in category 3 that are estimated to be shifted out of that category and into category 2 will be

$$p_c = (\hat{c} - c)/\hat{c} = 1 - (c/\hat{c}).$$

The *z* statistic in the last paragraph can be used for a two-tailed test of whether there the level of transfers is statistically significant, but it can also be used for one-tailed tests of whether there have been significant upward transfers or downward transfers.

This aggregated-data form of the model can be used to conduct a simple check for early evidence of age displacement in contexts where it seems likely, using unweighted numbers of cases in four successive age intervals.

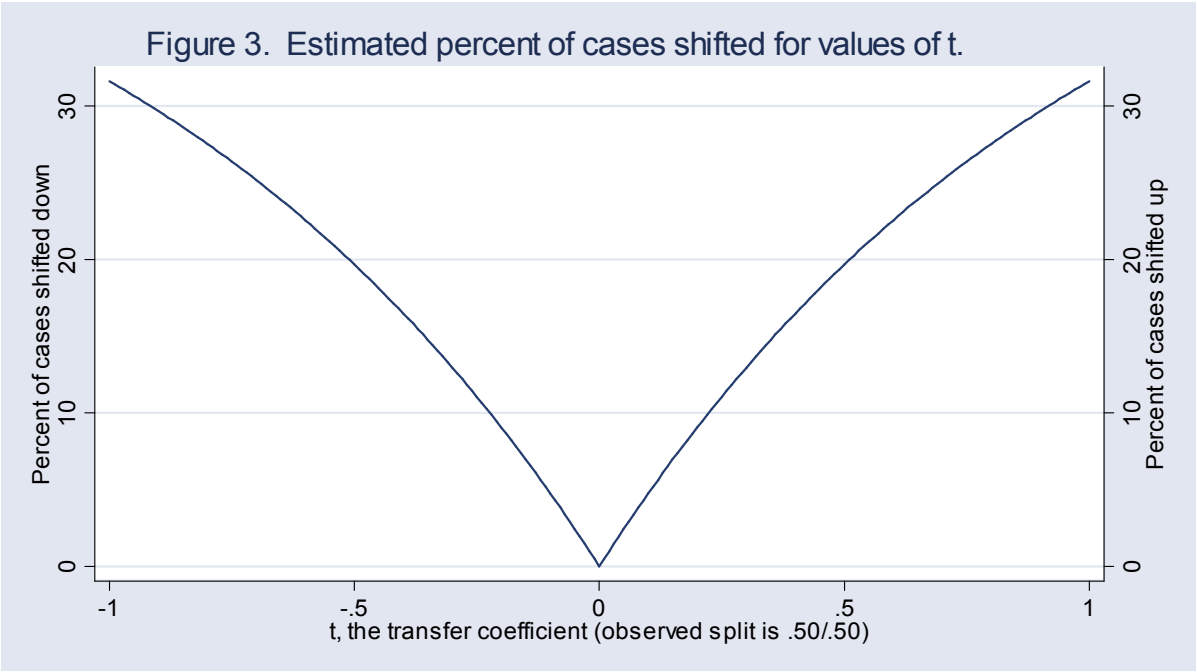


Figure 2 shows the relationship between the *t* coefficient and the proportions p_b and p_c, after multiplication by 100, when the observed values *b* and *c* are equal to each other. For example, if *t*=.5, we estimate that 20% of the true number of cases in the interval before the boundary have

been shifted upwards into the interval above the boundary. If t=-.5, then we estimate that 20% of the true number of cases in the interval above the boundary have been shifted downwards into the interval below the boundary. The conversion will be depend on the actual observed numbers in the two age intervals.

We will illustrate this approach by applying it to the females in the Zambia 2001/02 household survey (unweighted). The frequencies in age intervals 5-9 through 20-24 are a=3047, b=2684, c=2039, and d=1789, respectively. The age ratio for age 10-14 is 2684 / 3047=.88 and for age 15-19 is 2039 / 2684=.76, leading to AR_i-AR_o,=-12, suggesting some downward displacement. Our method gives $\ln(\hat{c}/\hat{b}) = [\ln(d/a)]/3 = -.1775$, leading to

$\hat{b} = (2684 + 2039)/1.8374 = 2570.5$ and

$\hat{c} = (2684 + 2039)(.8374/1.8374) = 2152.5.$

This suggests that a proportion $p_c = 1 - c/\hat{c} = 1 - 2039/2152.5 = .053$ or 5.3% of 15-19 year old women were misreported as 10-14. We get a transfer coefficient

$t = \ln(c/b) - [\ln(d/a)]/3 = \ln(2039/2684) - [\ln(1789/3047)]/3 = -.09735.$

The standard error of t is s=.03101 so $z = t/s = -3.14$.

All the coefficients described above for aggregate data can be replicated with a logit regression. The crucial step is the construction of two artificial variables. The first one, called x, distinguishes the first and fourth age intervals from the second and third. The second, called y, distinguishes the second interval in each pair from the first interval.

Table 4 defines a 2x2 table that gives x and y and the number of cases in each combination of the two. Specifically, if the interest is in age transfers across age 15, we construct a variable y which is 0 for age 5-14, 1 for age 15-24, and missing otherwise. We construct a variable x which is 0 for age 5-9 or 20-24, 1 for age 10-19, and missing otherwise.

In this table, the letters a, b, c, and d refer to the number of cases in four successive age intervals, sequenced as above. They are used as frequency weights (fweights in Stata) in a logit regression of y on x, which can be written as $\log it(\hat{y}) = b_0 + b_1 x$. (The Stata command would be logit y x [fweight=n], where n is the generic label for the frequencies.) The vector of coefficients will include intercept b_0 and slope b_1. In terms of the frequencies a, b, c, and d, the slope of the logit regression will be $b_0 = \ln(d/a)$ and the slope will be $b_1 = \ln(c/b) - \ln(d/a)$. The transfer coefficient can be calculated from the slope and intercept as

$t = \ln(b/c) - (1/3)\ln(d/a) = [\ln(b/c) - \ln(d/a)] + (2/3)\ln(d/a) = (2/3)b_0 + b_1.$

Table 4. Layout of the four successive age groups into a 2x2 table for logit regression approach to age tansfers. *a*: number of cases in first age group (e.g. 5-9), *b*: number of cases in second age group (e.g. 10-14), *c*: number of cases in third age group (e.g. 15-19), *d*: number of cases in fourth age group (e.g. 20-24).

		x = 0	*x* = 1
y	0	*a*	*b*
	1	*d*	*c*

The standard error of *t* is given by $s = \sqrt{(4/9)s_{00} + (4/3)s_{10} + s_{11}}$. The ratio $z=t/s$ will have a unit normal distribution under the null hypothesis that $t=0$ in the population. If *t* is positive, the estimated probability that a case in the second age interval will be transferred up to the third interval is

$$p_b = \frac{\exp(b_0 + b_1) - \exp(b_0/3)}{1 + \exp(b_0 + b_1)}.$$

If *t* is negative, the estimated probability that a case in the third interval will be transferred down to the second interval is

$$p_c = \frac{1 - \exp[(2/3)b_0 + b_1]}{1 + \exp(b_0 + b_1)}.$$

We note that the transfer coefficient can also be calculated by setting $x=3/2$ in the logit regression, giving $\log it(\hat{y}) = b_0 + 3b_1/2$, and then multiplying by 2/3. That is, $t = (2/3)\log it[\hat{y}(3/2)] = (2/3)b_0 + b_1$.

Application to individual-level data

The logit regression format can be easily extended to individual-level data. The variance-covariance matrix for the coefficients will include s_{00}, the estimated variance of the intercept, s_{11}, the estimated variance of the slope, and s_{10}, the estimated covariance of the intercept and slope.

One of the advantages of the individual-level format with logit regression is that it allows us to incorporate sampling weights and clustering. It also allows the inclusion of a covariate which may affect the amount of transfer. Suppose that a covariate *w*, and the interaction term *xw*, are added to the logit regression in the previous paragraph, so the model becomes

$\log it(\hat{y}) = b_0 + b_1 x + b_2 w + b_3 xw = (b_0 + b_2 w) + (b_1 + b_3 w)x$. (Of course, the coefficients with subscripts 0 and 1 will generally differ in value from the coefficients with the same subscripts in the previous model.) Then the measure of becomes a function of *w*, specifically $t(w) = (2/3)(b_0 + b_2 w) + (b_1 + b_3 w)$. The effect of a unit change in *w* on the transfer coefficient will be $\frac{d}{dz} t(w) = (2/3)b_2 + b_3$. (If *z* is binary and coded 0/1, then this will be the difference $t(1) - t(0)$.) The easiest way to assess whether *w* has a significant effect on the transfer rate is to divide the slope by its estimated standard error, $s = \sqrt{(4/9)s_{22} + (4/3)s_{32} + s_{33}}$, giving a unit normal test statistic of the null hypothesis that the slope is 0 in the population.

Some results from DHS surveys

The transfer procedure described here was used to check for evidence of transfers of women outside the age boundaries of the surveys of women, evidence of transfers of men outside the age boundaries of male surveys, and evidence of transfers of children outside the boundary of eligibility for health questions. Downward shifts of women or men and upward shifts of children were considered to be substantively significant if they exceeded 10%. Upward shifts of women or men were more common, and were considered to be substantively significant if they exceeded 20%.

The single survey with the strongest evidence of downward shifts of women was the 1990 survey of Nigeria. The numbers of women (de jure residence, weighted and rounded to the nearest integer) were as follows. The first column gives the observed frequencies and the second column gives the two fitted frequencies.

Age Interval	Observed Frequency	Fitted Frequency
5- 9	*a*=3974	*a*
10-14	*b*=3259	$\hat{b}$ =2832
15-19	*c*=1733	$\hat{c}$ =2159
20-24	*d*=1760	*d*

The proportion of "true" cases shifted downward from age 15-19 to age 10-14 is (1-1733/2159)=.197 or 19.7%. This is of course an estimate of net transfers; there were probably some upward and downward transfers that cancelled out. The transfer coefficient is t = -.3601, with standard error .0407, so the test statistic for the hypothesis of no downward transfers is z= -8.86. This is highly significant, but we caution against placing too much importance on evidence of statistical significance.

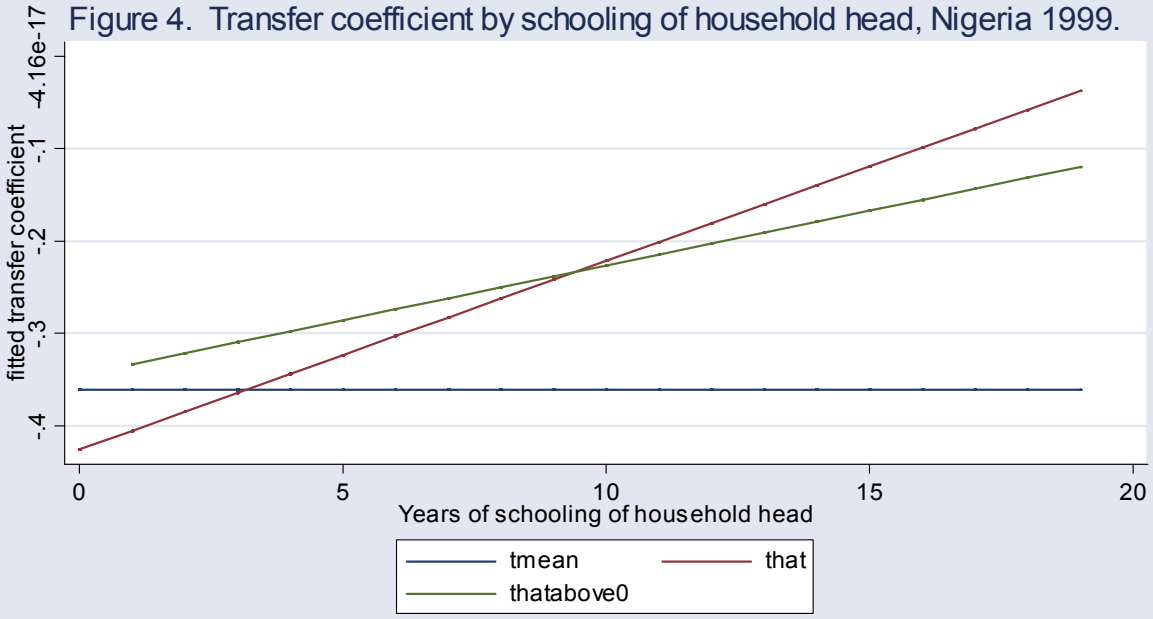
To see whether downward transfers could be related to education, we use the multivariate form of the model with *w*=completed years of schooling (hv108) of the household head. The distribution of this variable is very skewed. About 60% of household heads in the 1990 survey

of Nigeria had no schooling at all, and the mean was 3.2 years. The slope of the *t* coefficient with respect to education is .0204, with a standard error of .0088 and a *z* statistic of 2.31.

The horizontal line in Figure 3 is the overall mean value of the transfer coefficient. The figure also includes two lines that show the fitted value of the transfer coefficient as a function of years of schooling. They have the expected pattern: downward transfers are most likely when the houshold head has no schooling, and for the highest levels in the sample, it is negligible. Because the education distribution is so skewed, the figure shows the fitted values including households in which the household head has no formal schooling (the steeper line) and the fitted valus *omitting* such households (the line that is less steep).

In general one might expect a stronger relationship with the education of the household respondent, who provides the interviewer with the household data, than with the education of the household head. For this particular survey, it is not possible to link to that measure. It should also be noted that although schooling is a significant covariate of household transfers in this survey, that is not always the case. Finally, the level of displacement in this survey was so conspicuous that special efforts were made in the next survey of Nigeria, in 1999, to reduce it. The lower boundary of eligibility was set to 10, and later the women 10-14 were filtered out. This device seems to have been largely successful. It appears to have induced a high level of transfer from age 10-14 into age 5-9, but relatively few transfers from age 15-19 into age 10-14.

Figure 4. Transfer coefficient by schooling of household head, Nigeria 1999.



Another application of the procedure is to backward transfers of children's birthdates. The survey with the highest level of such transfers was the 1991 survey of Pakistan, a country that is well known for systematic upward transfers of children's ages in virtually every data collection procedure (see, for example, Pullum, 1990, and Pullum and Stokes, 1997). The window for the

extra health questions in this survey began in January 1986, so the four calendar years that straddle the boundary are 1984, 1985, 1986, and 1987. The numbers of births reported for those calendar years, weighted and rounded to the nearest integer, are given below. The progression across birth years is clearly much more plausible for the fitted frequencies than for the observed frequencies.

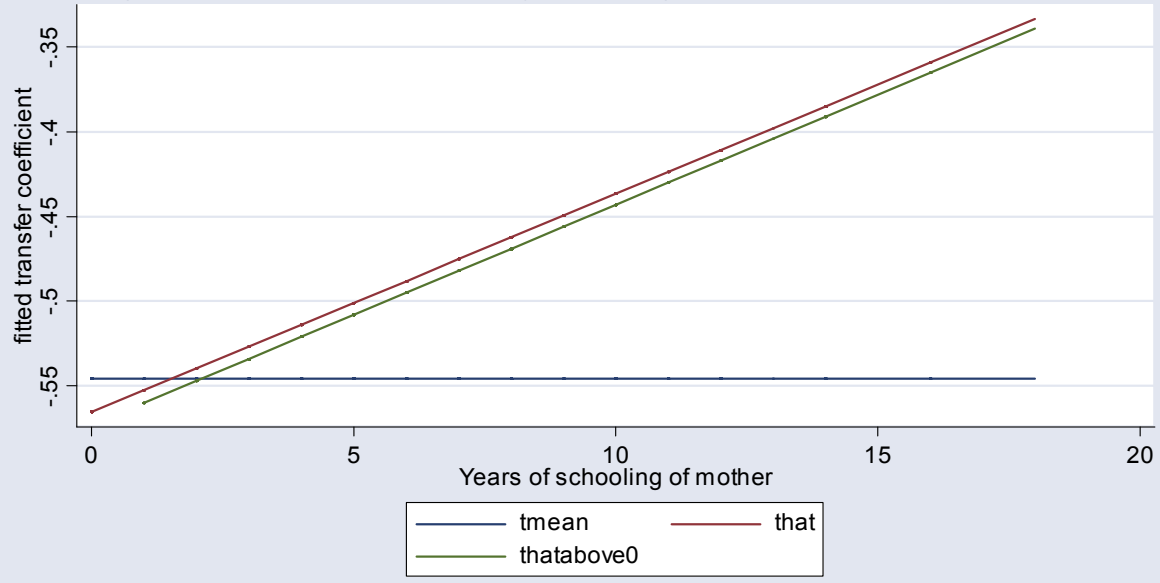
Time Interval	Observed Frequency	Fitted Frequency
1984	a=1790	a
1985	b=2033	$\hat{b}$ =1628
1986	c=1052	$\hat{c}$ =1457
1987	d=1285	d

The overall value of the transfer coefficient is -.5478, corresponding with backward transfers of 27.8% of the births that actually occurred in 1986 into 1985.

Figure 5 shows the linear relationship of the transfer function to the mother's years of schooling (v133). There is a fairly steep gradient by education, such that age transfers are less likely for women with more education, but even at the maximum level of education the estimated transfer coefficient would be .-35.

About 79% of the mothers of these children had no schooling at all. If the regression excludes those women, there is virtually no change in the regression line. Both lines are given in figure 4; the upper line includes women with no education and the lower line omits them. (The lines are so close that one might suspect a programming error, but they are in fact two distinct lines.)

Figure 5. Transfer coefficient by schooling of mother, Pakistan 1991.



7. Conclusions

This paper has attempted to move some longstanding demographic measures of the quality of age and date reporting into the modern era of statistical and individual-level data analysis. Only a few applications have been included, but these show how one might be able to develop a better understanding of the sources of incompleteness and misreporting.

There have been several references to standard errors, test statistics, and confidence intervals, but we recommend extreme caution in their use. For example, a test of whether there have been net transfers using the age ratio approach is simply a test of whether the reported age distribution is consistent with our model. At least some degree of deviation from the model would be expected even if there were no reporting error. The "true" age distribution is a consequence of a changing history of fertility and mortality (migration may play a role too) and we must try to avoid confounding genuine irregularities in the age distribution with misreporting. An appendix is included as an effort to assess the validity of this model. For evidence of transfers to be convincing, we advise a high standard such as 10% or 20% *as well as* nominal statistical significance.

The multivariate form of Myers' Index also depends on the validity of the adjustment or blending process, and the standard of a uniform distribution after blending. If Myers' Index is below 5% or so, there is probably little reason to be concerned. Tests related to incompleteness, on the other hand, do not depend on the "true" age distribution or birth dates. Tests related to heaping of age at death 12 months are probably robust with respect to the underlying "true" distribution of age at death because of a long reference period, such as ten years.

As mentioned, a natural extension of this work would be to some in-depth multivariate analyses of sources of misreporting. The DHS report by Pullum (2005) largely concerns the estimated levels of misreporting across virtually all DHS surveys, but only stratches the surface of an investigation into the sources. These methods could be used to look into variations by many characteristics of the respondents other than education. They also make it possible to look further into the role of the interviewer, using either fixed or random effects models. For example, how strong is the evidence that interviewers bear sole responsibility for age transfers? Does the probability that an interviewer will shift an age or birthdate depend on the respondent having low education? It seems likely that the interviewer will be more likely to shift an age if the respondent is less certain about the true age, suggesting the possibility of interaction effects, whereby combinations of *some* interviewers and *some* respondents are most likely to produce distortion.

Other characteristics of the interview could be important. It has been hypothesized that the actual time of the interview may induce some misreporting. For example, there may be a motivation to shorten the last interview of the day. Across the scores of interviews conducted by any given interviewer, the earliest interviews may differ in quality from the last interviews.

Finally, in assessing the consequences of misreporting, it is important to know the level at which it begins to impair the important summary measures of fertility, infant and child mortality, contraceptive use, child health, etc. Systematic omission of women age 45-49 from the survey of women, or of five-year-old children from the child health portion of the interview, may induce a bias in the estimates.

The DHS report (Pullum, 2005) includes comparisons between estimates of age-specific and total fertility rates for the same three-year reference period coming from successive surveys. For example, if surveys were conducted in the same country in 1995 and 2000, both surveys are used to estimate fertility rates during the calendar years 1992-94. Similar comparisons were done for infant mortality rates, but using a five-year reference period. Several countries show substantial differences between successive estimates, exceeding one child in the TFR for ages 15-44 or 50 points in the IMR. These discrepancies are probably mostly due to misreporting, especially transfers of young children, in one or both surveys, but the linkage is not clearly established. To the extent that both fertility and misreporting, or both infant mortality and misreporting, vary according to education or other covariates, misreporting will induce misleading estimates of the effect of education or other covariates on fertility. The multivariate approach in this paper may make it easier to model the consequences, as well as the sources of misreporting.

Appendix 1. Partial validation of the method to identify age transfers

In order to assess the validity of the modified age-ratio procedure to estimate age transfers, the version for aggregated data was applied to several five-year age distributions estimated for 1995 by the Population Division of the United Nations (United Nations, 2003). The year 1995 was selected partly because it was within the range of the DHS survey dates. It also preceded the 2000/2001 censuses that most countries had and was sufficiently in advance of the U.N. publication date (2003) that the tabulations are probably final estimates for 1995 and not contaminated with any projections. It is possible that in some cases the U.N. data were smoothed or interpolated with statistical and demographic methods, which would tend to make them more consistent with our model, but we have tried to avoid smoothed data.

Nine tabulated age distributions were selected, representing countries that showed high levels of transfers in the DHS report. Six distributions refer to women in African countries. The other three refer to men in Armenia, Kazakhstan, and Kyrgyzstan. The model was applied to five-year age groups in the range from 0-4 to 55-59, leading to an estimate of outward shifts for ages 5-9 through 50-54. The results are given in table A1. The percentages in column (3) are estimates of the percentage of "correct" cases that were shifted upward to the next interval, assuming that the age interval occupied position 2 in the four-interval model. The percentages in column (4) are estimates of the percentage of "correct" cases that were shifted downward to the next interval, assuming that the age interval occupied position 3 in the model. If the U.N. tabulations were correct, and if the assumptions of the model were correct, all of the percentages in columns (3) and (4) would be zero. Our interest is mainly in whether the model gives spurious evidence of shifts in the vicinity of ages 15 and 50, specifically in the estimates for age 15-19 in column (4) and for age 45-49 in column (3).

The results given in table A2 indicate that the model worked very well in the African countries. The spurious estimates of downward transfers from age 15-19 into 10-14 range from -0.3% to 2.5%. These are well below a threshold of 10% used in the report. The spurious evidence of upward transfers from 45-49 into 50-54 range from -4.0% to 2.8%, consistently well below the thresholds of 20% used in the report.

Considering that these results come from the African countries with greatest evidence of age transfers, we infer that the method was successful in that context. A somewhat different conclusion is reached for men in Armenia, Kazakhstan, and Kyrgyzstan, given in the last two panels of table A1. The percentages in columns (3) and (4) for these panels are generally larger in magnitude and more erratic than for the African women. These spurious estimates of transfers do not exceed the thresholds of 10% and 20% used in the report, but they would exceed thresholds of, say, 5% and 10%. There appear to be some genuine inconsistencies between the assumptions of the model and the actual age distributions for these countries.

Table A1. Estimated percentages shifted out (+) or in (-) when the four-category method is applied to UN age distributions (United Nations, 2003).

Column(1): Beginning of age interval
Column(2): Population (in thousands)
Column(3): Estimate when the age interval is the second in the sequence
Column(4): Estimate when the age interval is the third in the sequence

a. Burkina Faso females 1995.

(1)	(2)	(3)	(4)
0	1000	.	.
5	813	0.7	.
10	687	-0.3	-0.8
15	577	1.0	0.3
20	490	-0.1	-1.2
25	394	0.3	0.2
30	303	-2.2	-0.3
35	219	0.5	3.0
40	174	0.6	-0.6
45	147	0.0	-0.7
50	127	.	-0.0
55	112	.	.

b. Ghana females 1995.

(1)	(2)	(3)	(4)
0	1385	.	.
5	1261	2.1	.
10	1168	-2.1	-2.4
15	960	0.9	2.5
20	805	-1.4	-1.1
25	647	1.0	1.7
30	546	-0.1	-1.2
35	453	-0.4	0.1
40	371	0.4	
45	307	-0.1	-0.5
50	250	.	0.2
55	202	.	.

c. Kenya females 1995.

(1)	(2)	(3)	(4)
0	2214	.	.
5	2200	-1.1	.
10	1931	-0.4	1.2
15	1605	-0.1	0.4
20	1295	-0.9	0.1
25	1019	0.6	1.1
30	830	0.1	-0.8
35	672	1.2	-0.1
40	538	-0.7	-1.6
45	392	-4.0	1.0
50	274	.	5.2
55	243	.	.

d. Madagascar females 1995.

(1)	(2)	(3)	(4)
0	1238	.	.
5	999	0.7	.
10	846	0.2	-0.8
15	719	-0.2	-0.3
20	604	-0.2	0.2
25	507	0.2	
30	430	0.4	-0.2
35	364	1.1	-0.5
40	299	-3.8	-1.3
45	222	2.8	4.7
50	192	.	-3.5
55	160	.	.

e. Nigeria females 1995.

(1)	(2)	(3)	(4)
0	8843	.	.
5	7386	0.7	.
10	6288	-0.2	-0.8
15	5207	-0.4	0.3
20	4255	-0.0	0.4
25	3513	0.0	0.0
30	2938	0.2	-0.0
35	2487	1.5	-0.3
40	2099	-3.2	-1.8
45	1602	2.3	3.9
50	1371	.	-2.8
55	1128	.	.

f. Uganda females 1995.

(1)	(2)	(3)	(4)
0	2031	.	.
5	1610	0.6	.
10	1329	-0.1	-0.8
15	1095	0.8	0.1
20	904	-0.2	-1.0
25	708	-1.9	0.3
30	535	1.1	2.4
35	443	0.4	-1.3
40	374	0.4	-0.5
45	314	1.1	-0.4
50	256	.	-1.4
55	188	.	.

g. Armenia males 1995.

(1)	(2)	(3)	(4)
0	140	.	.
5	176	-3.0	.
10	167	-2.2	3.0
15	143	-0.9	2.4
20	127	-1.3	1.0
25	124	5.8	1.3
30	144	1.7	-5.6

35	141	-2.4	-1.8
40	105	-0.7	3.0
45	70	-17.3	1.0
50	44	.	19.0
55	79	.	.

h. Kazakhstan males 1995.

(1)	(2)	(3)	(4)
0	753	.	.
5	940	-7.4	.
10	810	1.1	7.4
15	750	1.2	-1.2
20	696	-6.5	-1.4
25	599	9.6	6.6
30	708	-3.3	-9.8
35	642	3.6	3.4
40	545	-7.8	-4.5
45	340	-5.5	10.4
50	271	.	6.2
55	393	.	.

i. Kyrgyzstan males 1995.

(1)	(2)	(3)	(4)
0	287	.	.
5	297	-2.5	.
10	259	-1.2	2.7
15	222	1.9	1.4
20	202	-3.3	-2.1
25	173	4.0	3.6
30	172	0.9	-4.4
35	154	-0.5	-1.1
40	117	-1.0	0.6
45	78	-14.1	1.4
50	49	.	16.5
55	74	.	.

Appendix 2. Illustrative computing routines in Stata

TO BE ADDED

Bibliography

Barclay, George. 1958. Techniques of Population Analysis. New York: John Wiley and Sons.

Chidambaram, V.C. and Thomas W. Pullum. 1981. Estimating fertility trends from retrospective birth histories: sensitivity to imputation of missing dates. *Population Studies 35*: 307-320.

Chidambaram, V.C. and Zeba A. Sathar. 1984. *Age and Date Reporting*. *WFS Comparative Studies No. 5*. Voorburg, Netherlands: International Statistical Institute.

Chidambaram, V.C., John G. Cleland, and Vijay Verma. 1980. *Some Aspects of WFS Data Quality: A Preliminary Assessment*. *WFS Comparative Studies No. 16*. Voorburg, Netherlands: International Statistical Institute.

Curtis, Sian L. 1995. *Assessment of the Quality of Data Used for Direct Estimatnion of Infant and Child Mortality in DHS-II Surveys. DHS Occasional Papers No. 3*. Calverton, Maryland: Macro International Inc.

Curtis, Sian L. and Fred Arnold. 1994. *An Evaluation of the Pakistan DHS Survey Based on the Reinterview Survey. DHS Occasional Papers No. 1*. Calverton, Maryland: Macro International Inc.

Demographic and Health Surveys. 1990. *An Assessment of DHS-I Data Quality. DHS Methodological Reports No. 1*. Columbia, Maryland: Institute for Resource Development/Macro Systems, Inc.

Ewbank, Douglas C. 1981. *Age Misreporting and Age-Selective Underenumeration: Sources, Patterns and Consequences for Demographic Analysis*. Committee on Population and Demography, Report No. 4. Washington, D.C.: National Academy Press.

Gage, Anastasia J. 1995. *An Assessment of the Quality of Data on Age at First Union, First Birth, and First Sexual Intercourse for Phase II of the Demographic and Health Surveys Program. DHS Occasional Papers No. 4*. Calverton, Maryland: Macro International Inc.

Goldman, Noreen, Ansley J. Coale, and Maxine Weinstein. 1979. *The Quality of Data in the Nepal Fertility Survey*. WFS Scientific Reports, No. 6. Voorburg, Netherlands: International Statistical Institute.

Goldman, Noreen, Shea Oscar Rutstein, and Susheela Singh. 1985. *Assessment of the Quality of Data in 41 WFS Surveys: A Comparative Approach*. WFS Comparative Studies, No. 44. Voorburg, Netherlands: International Statistical Institute.

Hobbs. Frank. 2004. Age and Sex Composition. In Jacob S. Siegel and David A. Swanson (eds.), *The Methods and Materials of Demography*, second edition. Elsevier Academic Press, pp. 125-173.

Marckwardt, Albert M. and Shea Oscar Rutstein. 1996. *Accuracy of DHS-II Demographic Data: Gains and Losses in Comparison with Earlier Surveys. DHS Working Papers No. 19*. Calverton, Maryland: Macro International Inc.

Mason, Karen O. and L.G. Cope. 1987. Sources of age and date-of-birth misreporting in the 1900 U.S. census. *Demography 24*: 563-73.

Nag, Moni H., E.G. Stockwell, and L.M. Snavely. 1973. Digit preference and avoidance in the age statistics of some recent African censuses: some patterns and correlates. *International Statistical Review 41*:165-74.

Potter, Joseph E. 1977. Problems in using birth history analysis to estimate trends in fertility. *Population Studies 31*:335-64.

Pullum, Thomas W., N. Ozsever, and T. Harpham. 1984. *An Assessment of the Machine Editing Policies of the World Fertility Survey*. *Scientific Report Serie, No. 54*. World Fertility Survey (International Statistical Institute). 39pp.

Pullum, Thomas W., T. Harpham, and N. Ozsever. 1986. The machine editing of large sample surveys: the experience of the World Fertility Survey. *International Statistical Review 54*: 311-326.

Pullum, Thomas W. Analytic methodology. 1987. In J. Cleland and C. Scott (eds.), *The World Fertility Survey: An Assessment of its Contribution*. Oxford University Press, pp. 644-676.

Pullum, Thomas W. 1988 *An Approach to the Reconciliation of Demographic Survey Data from the Philippines. Occasional Paper No. 2.* Population Technical Assistance Project (International Science and Technology Institute). 29 pp.

Pullum, Thomas W. 1990. Statistical methods to adjust for date and age misreporting to improve estimates of vital rates in Pakistan. *Statistics in Medicine 10*:191-200.

Pullum, Thomas 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. Paper presented at 2005 Annual Meetings of the Population Association of America.

Pullum, Thomas W. and S. Lynne Stokes. 1997. Identifying and adjusting for recall error, with application to fertility surveys. In Lars Lyberg and Paul Biemer, et al. (eds.), *Survey Measurement and Process Quality*. John Wiley and Sons, pp. 711-732.

Rutstein, Shea O., and George T. Bicego. 1990. Assessment of the quality of data used to ascertain eligibility and age in the Demographic and Health Surveys. In *An Assessment of DHS-I Data Quality. DHS Methodological Reports, No. 1*. Columbia, Maryland: Institute for Resource Development/Macro Systems, Inc.

Shryock, Henry S., Jacob S. Siegel, and Associates. 1971. *The Methods and Materials of Demography*. U.S. Bureau of the Census, U.S. Government Printing Office.

Stanton, Cynthia, Noureddine Abderrahim, and Kenneth Hill. 1997. *DHS Maternal Mortality Indicators: An Assessment of Data Quality and Implications for Data Use. DHS Analytical Report No. 4*. Calverton, Maryland: Macro International Inc.

Swanson, David, and Jacob S. Siegel (eds.) 2004. *The Methods and Materials of Demography*. Academic Press.

United Nations. 1967. *Manual II: Methods of Appraisal of Quality of Basic Data for Population Estimates. Population Studies, No. 23*. New York: Department of Economic and Social Affairs.

United Nations. 1987. *A Comparative Evaluation of Data Quality in Thirty-Eight World Fertility Surveys*. New York: Department of International Economic and Social Affairs.

United Nations. 2003. *World Population Prospects: Tthe 2002 Revision. Volume II: Sex and Age Distributions of Populations*. New York: Department of Economic and Social Affairs.

R.J.A. Little and T.W. Pullum. The general linear model and direct standardization: a comparison. *Sociological Methods and Research 7* (1979): 475-501.

T.W. Pullum. Analytic methodology. In J. Cleland and C. Scott (eds.), *The World Fertility Survey: An Assessment of its Contribution*. Oxford University Press, 1987: 644-676.