Omissions of Births in DHS Birth Histories in Sub-Saharan Africa:
Measurement and Determinants

DRAFT – PAA, Washington DC, March 31-April 2, 2011

Bruno SCHOUMAKER, Centre de recherche en démographie et sociétés, Université catholique de Louvain

1. Background and objectives

Birth histories have been collected for more than 30 years in developing countries – notably in the WFS and DHS programs - and have become a major source for estimating fertility levels and trends. With close to 100 surveys conducted in sub-Saharan Africa since the mid 1980s, the DHS program has allowed tremendous progress in the knowledge of fertility levels, trends, and determinants in that region.

Despite the usefulness of birth histories, they are subject to several data quality problems (Arnold, 1990; Blacker, 1994; Chidambaram et al., 1980; Choi and Sudhinaraset, 2010; Goldman et al., 1985; Kreyenfeld et al., 2010; Pullum, 2006). One of the most commonly mentioned problems of DHS birth histories is the displacement of births (Arnold, 1990; IGCME, 2009; Pullum, 2006). This is linked to the design of the DHS questionnaire, and to the fact that some interviewers can change the birth dates of certain children to avoid having to administer the lengthy health modules in the DHS[1]. This problem was identified early in the DHS program (Arnold, 1990), and is still a serious issue in sub-Saharan Africa (Pullum, 2006). This phenomenon leads to underestimating recent fertility, and to overestimate past levels of fertility[2].

Omission of births is another limitation of birth histories (Blacker, 1994; Merli and Raftery, 2000). Early evaluations of data quality of birth histories – notably in the World Fertility Surveys - led to the conclusion that omissions of births were "most likely to affect fertility at young ages for older cohorts" (Brass and Rashad, 1980, cited in United Nations, 1987b). However, omission of <u>recent</u> births is another serious issue which, like displacement of births, may be linked to the lengthy health modules. Severe omissions of recent births

[1] Questions on health of children are limited to births that occurred a few years before the survey. The cut-off date for the health module is often January five years before the survey; but the length of the reference period varies across surveys.

[2] However, the impact of this problem can be reduced by choosing appropriate periods for computing fertility rates. (Sullivan, REF)

have been suspected in several DHS in sub-Saharan Africa. For instance, in the 1999 Nigeria DHS, "[…] omission of births in the three-year period immediately prior to the survey […resulted] in an underestimate of current fertility of about 16-17 percent" (National Population Commission – Nigeria, 2000, p.36). As this example shows, omissions may lead to severe underestimates of fertility levels. Differential omissions across surveys can also have serious consequences on the measurement of fertility trends (Schoumaker, 2008; Chidambaram, 1980).

Surprisingly, little work has been done to detect and quantify omissions of births in DHS and to identify factors that account for variations of omissions across surveys. In this paper, we investigate omissions of recent births in DHS in sub-Saharan Africa in a systematic way, and identify some factors that influence the degree of omissions. This paper is organized around three objectives:

1. The first objective is to propose a method to measure omissions of recent births, based on the comparison of birth histories from consecutive surveys.
2. The method is used to measure omissions of births in 52 DHS in sub-Saharan Africa. Omissions of births are measured at the country level and for several subgroups of births (gender, survival status) or women (education).
3. Correlates of omissions are then identified using regression models. The dependant variable is the degree of relative omissions in each of the 52 surveys. Independent variables include questionnaire design variables (length of questionnaire, length of reference period), variables related to the size of the survey (number of interviewers, interviews per interviewer) and characteristics of the population (level of education).

We show that omissions of recent births are widespread in Sub-Saharan Africa, leading to significant underestimates of fertility. Omissions are more widespread among the less educated, and among deceased children. Three major factors explaining omissions are identified: length of the questionnaire, length of the reference period of the health module, and level of education of the population.

2. Measuring omissions of recent births: a brief review of approaches

Detecting omissions from a single survey

The measurement of omissions of recent births is, by nature, not a straightforward task. Several approaches are typically used, all of which have shortcomings[3]. The sex ratio at birth and the comparisons of levels of neonatal mortality compared to infant mortality (Arnold, 1990) are classical indicators to identify omissions of births from a single survey. The former suggests differential omissions by gender (Arnold, 1990; Narasimhan et al. 1997); the latter indicates possible omissions of children deceased early after their births (neonatal deaths). These indicators measure selective omissions. They can be useful to detect serious omissions, but are not capable of detecting non-selective omissions, and are not useful to rule out omissions altogether (Sullivan et al., 1990).

The P/F method, originally developed by Brass, has also been used to detect underreporting of births from a single survey. For instance, Cleland et al. (1994) used it to evaluate data quality of fertility in a large number of surveys and censuses in sub-Saharan Africa. As noted by the authors, the P/F ratio has some limitations to detect underreporting of births, as its interpretation is complex when fertility changes[4]. Moreover, birth displacement and omissions are not distinguished. Despite the limitation of the method, their study suggested that underreporting of recent births was quite frequent in sub-Saharan Africa.

Detecting omissions from multiple sources

The comparison of several data sources is another possible approach to detect and measure omissions of births. A reliable civil registration system would make it an easy matter (Sullivan 2008). However, this option is not available in most sub-Saharan African countries, where civil registration systems are unreliable (Tabutin and Schoumaker, 2004). Census data are also not an option to evaluate omissions in a systematic way. Fertility is often underestimated in censuses (Ortega, 2008), and the dates of the censuses may not correspond to fertility estimates from DHS. Other surveys, like MICS surveys, may also give hints about possible omissions, but these surveys are also likely to be affected by data quality problems.

---

[3] In this paper, we do not discuss omission of births in the distant past. The typical method for detecting omissions at young ages among older cohorts consists in inspecting “increments in mean parity across age groups” (Chidambaram, 1980, p.20).

[4] The general idea is to compare period fertility (F) cumulated to age X to parity (P) the same age. If fertility is constant, the ratio P/F should be close to one. A ratio greater than one suggests underreporting of births in the covered period (because of displacements or omissions). However, a ratio greater than one may also be consistent with a fertility decline (Cleland et al. 1994, p.8)

When two or more DHS surveys are available, a powerful approach to evaluate data quality of birth histories consists in comparing the level of fertility in one survey (called the 'first survey') with the level fertility for the same period estimated from a subsequent survey (called the 'second survey'). This general approach has been used in several contexts (Arnold, 1990; Pullum, 2006; Schoumaker, 2008). Several options can be envisaged. The simplest form consists in comparing TFRs for the same period from successive surveys. Using that method, Pullum (2006) showed that discrepancies between TFR (15-44) from successive surveys were often very large, notably in Sub-Saharan Africa. Again, a limitation is that it is not possible to attribute the differences across surveys to omissions or displacements[5].

More complex approaches rely on the reconstruction of fertility trends. A graphical representation of annual TFRs from successive surveys provides a useful way of detecting omissions and displacements (Arnold, 1990; Schoumaker, 2008[6]). However, this approach does not allow quantifying omissions in a straightforward way and it may be difficult to detect omissions that are not severe. In this paper, we propose a method based on the reconstruction of fertility trends using Poisson regression to detect and measure omissions in a systematic way.

3. Discrepancies across surveys: comparisons of TFRs

We first use similar approach as the one used by Pullum (2006) to illustrate differences across surveys. It consists in comparing the recent fertility (TFR) in one survey ('first survey') with the TFR for the same period computed using data from a subsequent survey ('second survey')[7]. To get comparable TFRs, the same periods are used for both surveys, and are defined as the three years prior to the mean date of interview (month) in the first survey[8]. TFRs are computed between age 15 and 44, as there are usually no or few women above 45 in the second survey to compute fertility rates in the 3 years preceding the first survey. The major interest of this method is that it is relatively simple to implement, and it gives a good indication of the discrepancies across surveys.

---

5 Interestingly, Pullum (2006) seems to attribute all the differences to birth displacements, and does not mention omissions. One of our conclusions is that omissions are the major source of discrepancies between TFRs.

6 See Retherford (1987) for a similar approach with the own children method, and Sullivan (1998) for child mortality.

7 52 pairs of successive surveys in the same country are used. The 'second survey' for one comparison can be the 'first survey' for another comparison.

8 This is almost the same approach as the one used in DHS reports, where fertility is measured for the three years prior to the survey month, which may vary across respondents. It is slightly different from Pullum's approach, where the TFR is computed for the "three calendar years prior to the median year of interview in the first survey" (Pullum, 2006).

Figure 1 compares the TFR estimated in the first survey (Y-axis) with the TFR computed in the second survey (X-axis). It shows that recent fertility is virtually always lower in the first survey than in the second survey (in 49 out of 52 surveys), and differences are statistically significant in a large number of surveys[9]. The absolute differences range from close to zero (e.g. 1988 and 1993 surveys in Ghana) to more than 2 children in Guinea (1999), and relative differences peak at 30% (1999 Guinea survey). On average, differences amount to 0.82 children (12%).

Data

The analyses rely on data from the Demographic and Health Surveys conducted in Sub-Saharan Africa since the mid 1980s. We retain all the countries of Sub-Saharan Africa where at least two comparable surveys have been conducted, with data published on the STATcompiler website, and for which data files are available[10]. Data from 75 surveys are used to measure omissions in 52 surveys from 23 countries.

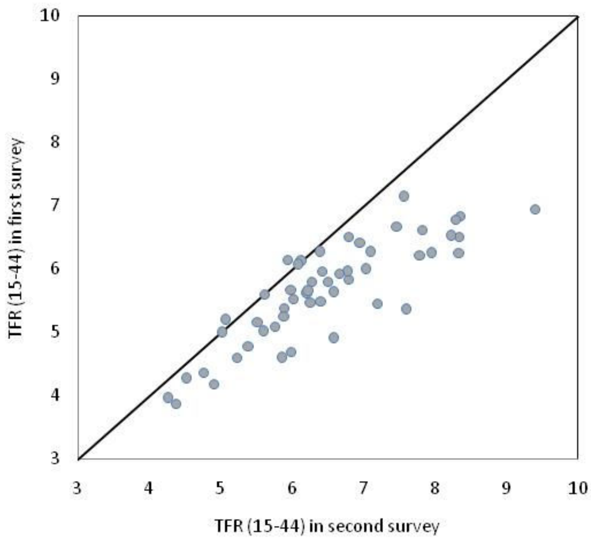
Both published data and individual data files are used. Published data are taken from the STATcompiler website (www.measuredhs.com) and from DHS reports, and concern some explanatory variables used in the last part of the paper. Individual data files are used to compute TFRs and to measure omissions and displacements. Some explanatory variables were also computed from individual data files.

These results show that there are important biases in fertility measurement in many DHS in sub-Saharan Africa. Some Sahelian countries (Burkina Faso, Mali, Niger, and Guinea) are particularly affected, as well as Mozambique and Cameroon. Even in countries where differences are relatively small, they can vary significantly from one survey to another. For instance, in Ghana, the difference was small in the first two surveys (close to zero), increased in the 1998 survey (0.41 child), and decreased somewhat in the 2003 survey (0.25 child). This issue is thus important not only for measuring fertility levels, but also for analyzing fertility trends.

[9] Differences are statistically significant at the 5% level in 37 out of 52 surveys. Standard errors for TFRs were computed using jackknife and take into account sample design (stratification and clustering). Two-tailed tests were performed.

[10] Only TFRs from 'Standard DHS' are published on the STATcompiler website. Fertility artes from surveys such as the AIS (AIDS indicator surveys) are not used in this paper. Surveys from Eritrea are not used because the data files are not available. Liberia is not included either, becaue the time interval between the two surveys (1986 and 2007) is too long to be meaningful in this study. Lesotho's surveys are not included, as the second Lesotho survey was released a few days before the first version of this paper.

Figure 1: Differences of TFRs (15-44) across DHS surveys in Sub-Saharan Africa



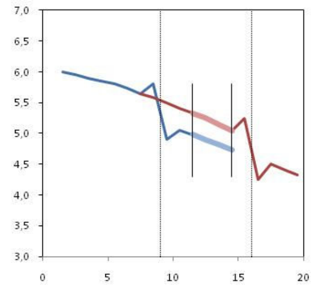
Sources of discrepancies

The differences between the TFRs in the first survey and in the second survey can result from several factors. They can reflect either an underestimation of fertility in the first survey, an overestimation of fertility for the same period in the second survey, or both. Underestimation and overestimation of fertility can themselves occur for several reasons. Underestimation of recent fertility in the first survey can potentially be due to birth displacement and omissions. Overestimation in the second survey can also be due to the displacements of births from earlier periods. Finally, differences in sample implementation can also lead to overestimation or underestimation of fertility in both surveys. For instance, if educated women were accidentally oversampled in the first survey, fertility would be biased downward.

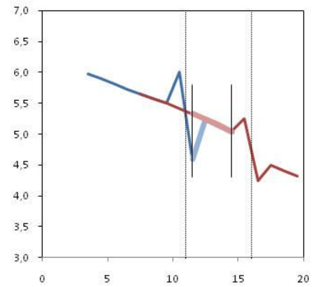
Figures 2.1 to 2.8 illustrate some situations that can lead to fertility differences between the first and the second survey. The estimation period (3 years) is delimited by two vertical solid lines in the middle of the graph, and fertility trends are thicker between these lines. The cut off year of the health module in each survey is represented by a vertical dotted line. We assume, for the sake of clarity, that births are displaced from the year following the cut-off date to the year just before (as suggested by evidence from other research; Arnold 1990; Pullum, 2006). We also represent omissions as constant for the period after the cut-off dates, and consider that differences in sample implementation lead to broadly constant differences between surveys. The first four figures represent situations where only one problem affects fertility level[11]. The next figures illustrate combinations of two or more data quality problems.

Figure 2.1 to 2.8 : Illustration of sources of discrepancies of TFRs across surveys

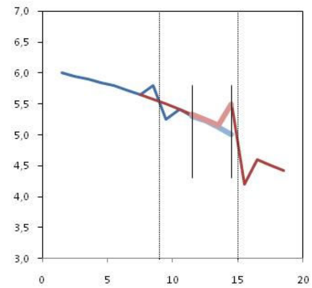
1) Omissions in first survey only



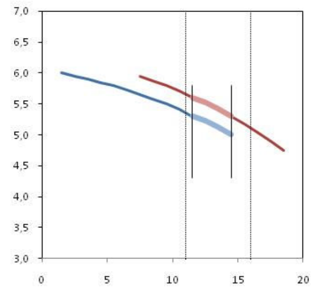
2) Displacements in first survey



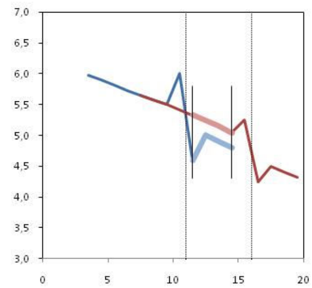
3) Displacements in second survey



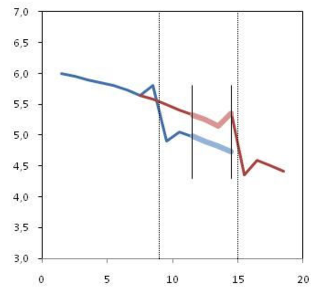
4) Sample implementation in first and/or second survey



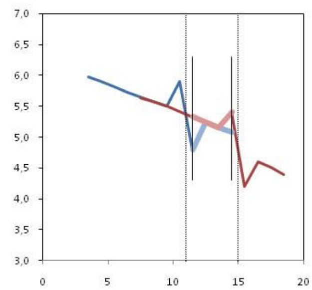
5) Omissions and displacements in first survey



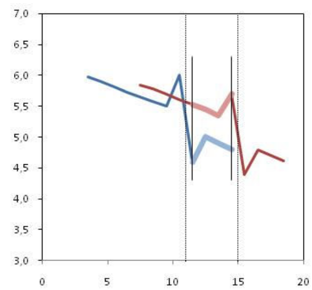
6) Omissions in first survey and displacement in second survey



7) Displacement in first survey and in second survey



8) Omissions and displacements in first survey, displacement in second survey, and sample implementation in first and/or second survey



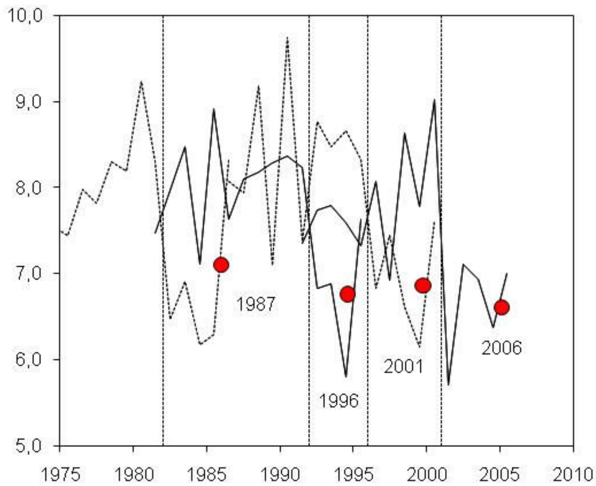
[11] This does not mean that only one problem is present in the data, but it may be present and not influence fertility level.

In Figure 2.1, the difference between the first and the second survey is due to omissions of recent births in the first survey. These surveys are also affected by displacements of births, but they do not influence fertility in the 3 years preceding the survey. Figure 2.2 represents a situation with birth displacements in the first survey. Such a situation is likely to occur if the reference period for the health module is close to three years. In that case, recent fertility would be underestimated. In the third example (Figure 2.3), fertility in the second survey is overestimated because some births are displaced backward in time. This is more likely to occur if the surveys are close to each other. In the fourth example (Figure 2.4), fertility trend are similar in both surveys, but differences in sample implementation lead to differences in fertility levels. Figures 2.5 to 2.8 illustrate combinations of several problems. Figure 2.5 combines omissions and displacements in the first survey. Figure 2.6 combines omissions in the first survey and displacements in the second survey. In Figure 2.7, both displacements of births in the first and in the second survey lead to a difference in fertility between surveys. Finally, figure 2.8 combines four types of problems: omissions and displacement in the first survey, displacement in the second survey, and differences in sample implementation.

4. Measuring omissions through modeling fertility trends

The comparison of annual values of TFRs from consecutive surveys provides a useful way to evaluate the presence of omissions and displacements. Figure 3 shows annual total fertility rates (15-49) reconstructed for the 15 years preceding each of the four DHS in Mali (1987, 1996, 2001, 2006). Red dots indicate TFRs estimated from the three years preceding the survey (published by Macro International). The vertical dotted lines represent the cut-off year of the health module in each survey. This figure illustrates the sharp drop in the TFR just at the cut-off year. It also shows that fertility after the cut-off year is lower higher than fertility for the same period estimated from the next survey. These two elements suggest that omissions are significant. It is, however, difficult from this figure to quantify it, and to separate omissions from displacements.

Figure 3: Reconstruction of TFRs (15-49) for the 15 years preceding the four DHS in Mali.



Poisson regression with person period data

The method we use to measure omissions relies on reconstructing fertility trends using Poisson regression with restricted cubic splines, and including dummy variables in the model to capture omissions and displacements of births[12]. The method is briefly presented below (for more details, see Schoumaker, 2010).

First, all the surveys for the same country are pooled together, and birth histories are transformed into a person period data set (Schoumaker, 2004). Each line in the data set corresponds to a period during which the 5-year age groups and the calendar years are constant. 5-year age groups (dummy variables) and a spline fuction of time are used as independent variables in a Poisson regression. The dependant variable of the regression is the number of births in each period, and the varying length of the periods is controlled as an offset. Restricted cubic splines (RCS) are used to model fertility trends. The basic principle of RCS is to fit piecewise polynomial functions constrained to join at knots (Andersen, 2009). To fit restricted cubic splines with K knots, K-1 variables (functions of time periods) are created, and introduced as independent variables in the Poisson regression[13].

[12] Differences in sample implementation could in principle also be taken into account, by including a dummy variable for each survey. We have shown elsewhere (Schoumaker, 2008) that differences in sample implementation are likely to have small impacts. Moreover, the sample implementation parameter may be biased by omissions in the past. For these reasons, we do not include it in the model.

[13] In this paper, knots are located every five years. As a result, 4 or 5 knots are used in most countries. This is sufficient to reach a good compromise between flexibility and rigidity (Harell, 2001). The location of knots every five years is done backward, starting from the last knot which is located on the year just before the cut-off year of the health module in the latest survey. The mkspline command in Stata creates automatically these variables after the number and the location of knots have been defined (StataCorp, 2007).

Next, a series of dummy variables is created to capture birth displacements and omissions. For each survey included in the pooled data set, three dummy variables are computed. This illustrated with in Figure 4 and Table 2 for the case of Mali (four surveys, subscripts indicate the survey number). The first variable (O_1), is a variable capturing omissions after the cut-off year of the health module in the first survey (1987). The health module started in January 1982, so that the O_1 variable is equal to 1 from 1982 to 1986 (the last year of the survey is dropped), and to 0 for all the other years. The second variable (DB_1) is a dummy variable capturing displacements of births to the year before the cut-off year. It is equal to zero, except for the year just before the cut-off year of the health module, where it is equal to 1. The third variable (DA_1) is a dummy variable capturing displacements of births from the cut-off year. It is equal to zero, except for the cut-off year of the health module, where it is equal to 1. For the 1987 survey, The DB variable is equal to 1 in 1981, and the DA variable is equal to 1 in 1982. A similar coding procedure is used for the 1996, 2001 and 2006 surveys.

Figure 4: Computation of dummy variables to measure omissions and displacement of births

$O_{1t} = 1$ if $t \geq COY_1$ and $survey = 1$; otherwise $O_{1t} = 0$
$O_{2t} = 1$ if $t \geq COY_2$ and $survey = 2$; otherwise $O_{2t} = 0$
$O_{3t} = 1$ if $t \geq COY_3$ and $survey = 3$; otherwise $O_{3t} = 0$
$O_{4t} = 1$ if $t \geq COY_4$ and $survey = 4$; otherwise $O_{4t} = 0$

$DA_{1t} = 1$ if $t = COY_1$ and $survey = 1$; otherwise $DA_{1t} = 0$
$DA_{2t} = 1$ if $t = COY_1$ and $survey = 2$; otherwise $DA_{2t} = 0$
$DA_{3t} = 1$ if $t = COY_1$ and $survey = 3$; otherwise $DA_{3t} = 0$
$DA_{4t} = 1$ if $t = COY_1$ and $survey = 4$; otherwise $DA_{4t} = 0$

$DB_{1t} = 1$ if $t = COY_1 - 1$ and $survey = 1$; otherwise $DB_{1t} = 0$
$DB_{2t} = 1$ if $t = COY_1 - 1$ and $survey = 2$; otherwise $DB_{2t} = 0$
$DB_{3t} = 1$ if $t = COY_1 - 1$ and $survey = 3$; otherwise $DB_{3t} = 0$
$DB_{4t} = 1$ if $t = COY_1 - 1$ and $survey = 4$; otherwise $DB_{4t} = 0$

COY=Cut-off year of the health module
t=year

These variables are included in the regression model. However, the data are dropped from the cut-off year of the last survey (2001) – because omissions cannot be estimated by comparison with another syrvey. As a result, the O4 and DA4 are all equal to zero, and not included in the model. In contrast, the DB4 variable, measuring a possible overestimation of fertility in the fourth survey in 2000 is included in the model.

[Eq. 1] describes the Poisson regression model.

$$\log(\mu_i) = \log(t_i) + f(age) + g(time) + \beta_1.O_1 + \beta_2.O_2 + \beta_3.O_3 + \beta_4.DB_1 + \ldots \qquad \text{[Eq. 1]}$$

μ_i is the expected number of children born in each time segment, t_i is the length of the time segment (exposure), f(age) is a function of age, and g(time) is the spline function of the calendar time[14]. O1, O2, O3 are the dummy variables for omissions, and the coefficients β_1, β_2 and β_3 measure omissions in the first three surveys. The coefficients of the O variables are expected to be negative, as they measure the ratio of fertility in periods affected by omissions in a survey compared to fertility levels in the same period in the next survey(s) not affected by omissions. The coefficients of the DB variables are expected to be positive, since displacement of births will increase fertility just before the cut off year, and the coefficients of the DA variables are expected to be negative, reflecting the decrease of fertility in the cut off year.

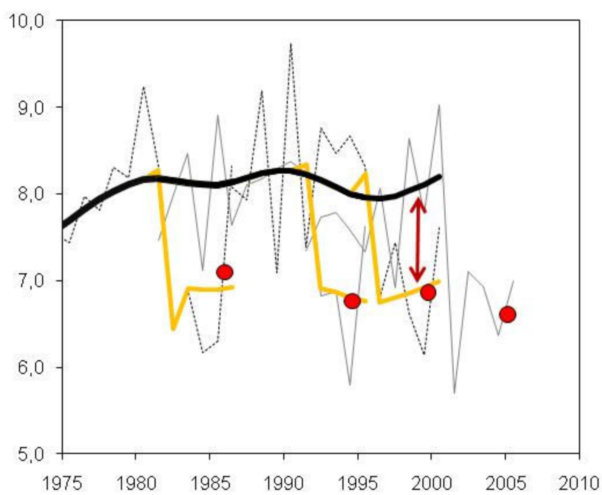
Table 2: Illustration of dummy variable coding for the measurement of omissions and displacement of births (grey cells indicate that data are not included in the model).

	First survey (1987) Cut-off year : 1982			Second survey (1996) Cut-off year : 1992			Third survey (2001) Cut-off year : 1996			Third survey (2006) Cut-off year : 2001		
Year	O1	DB1	DA1	O2	DB2	DA2	O3	DB3	DA3	O4	DB4	DA4
1977	0	0	0	0	0	0	0	0	0	0	0	0
1978	0	0	0	0	0	0	0	0	0	0	0	0
1979	0	0	0	0	0	0	0	0	0	0	0	0
1980	0	0	0	0	0	0	0	0	0	0	0	0
1981	0	1	0	0	0	0	0	0	0	0	0	0
1982	1	0	1	0	0	0	0	0	0	0	0	0
1983	1	0	0	0	0	0	0	0	0	0	0	0
1984	1	0	0	0	0	0	0	0	0	0	0	0
1985	1	0	0	0	0	0	0	0	0	0	0	0
1986	1	0	0	0	0	0	0	0	0	0	0	0
1987	0	0	0	0	0	0	0	0	0	0	0	0
1988	0	0	0	0	0	0	0	0	0	0	0	0
1989	0	0	0	0	0	0	0	0	0	0	0	0
1990	0	0	0	0	0	0	0	0	0	0	0	0
1991	0	0	0	0	1	0	0	0	0	0	0	0
1992	0	0	0	1	0	1	0	0	0	0	0	0
1993	0	0	0	1	0	0	0	0	0	0	0	0
1994	0	0	0	1	0	0	0	0	0	0	0	0
1995	0	0	0	1	0	0	0	1	0	0	0	0
1996	0	0	0	0	0	0	1	0	1	0	0	0
1997	0	0	0	0	0	0	1	0	0	0	0	0
1998	0	0	0	0	0	0	1	0	0	0	0	0
1999	0	0	0	0	0	0	1	0	0	0	0	0
2000	0	0	0	0	0	0	1	0	0	0	1	0
2001	0	0	0	0	0	0	0	0	0	1	0	1
2002	0	0	0	0	0	0	0	0	0	1	0	0
2003	0	0	0	0	0	0	0	0	0	1	0	0
2004	0	0	0	0	0	0	0	0	0	1	0	0
2005	0	0	0	0	0	0	0	0	0	1	0	0

[14] This approach assumes that there is no interaction between the age effect and the period effect (time); i.e. that the shape of the age-specific fertility rates is constant over time. Although the assumption does not strictly hold, simulations indicate that violating this assumption does not have strong influences on fertility levels and trends

Figure 5 illustrates the method in a graphic way in the case of Mali. As in Figure 3, fine dotted lines and solid lines represent total fertility rates (15-49) estimated from each survey separately. The black curve shows the reconstructed fertility trend with Poisson regression using the four pooled surveys and restricted cubic splines. It is computed by predicting total fertility rates from the regression coefficients, but excluding omissions and displacements from the prediction equation. In other words, this trend is corrected for omissions and displacements. The orange lines indicate the predicted fertility levels for periods affected by omissions and displacements (coefficients of omissions and displacements are included in the prediction equation for each survey). This figure shows for instance that recent fertility estimated in the 2001 survey is much lower (15%, around 1.2 children) than fertility for the same period measured with retrospective data from the 2006 survey. This relative difference (interpreted as relative omissions) is simply obtained from the regression coefficients of the dummy variables included in the models[15]. In the case of Mali, omissions are the major issue; in contrast, birth displacements are not pronounced.

Figure 5: Reconstruction of fertility trends (TFR) in Mali and estimation of birth omissions (source of data: DHS 1987, 1996, 2001 and 2006).



The Poisson regression model is fitted in each of the 23 countries, and coefficients of omissions and displacements are estimated in 52 surveys. Analyses are first performed at the country level. Next, separate analyses are done by level of education, and by gender and survival status of children. Analyses by education simply consist in doing the analyses separately by level of education of women (no schooling; some schooling). Analyses by

[15] It is equal to exp(β)-1, where β is the coefficient of the omission variable.

gender and survival status of children consist in recoding births. For instance, for analyses on girls, births of boys are recoded as zero (fertility rates only include births of girls as events, but all women years in the denominator). A similar approach is used for estimating omissions by survival status, depending on the survival status of children at 6 months[16].

5. How widespread are omissions of births in sub-Saharan Africa?

Table 3 shows that omission of recent births in DHS is a widespread and systematic phenomenon in sub-Saharan Africa. The omission parameter is negative in 51 of the 52 surveys, and is statistically significantly negative in 40 of these surveys[17]. On average, relative omissions amount to 9%, and range from 23% in the 1987 Mali DHS to close to zero in several surveys (Kenya 1988, Ghana 1988, Zimbabwe 1988)[18]. Again, some countries and surveys are particularly affected (Guinea, Mozambique, Mali…).

Table 3: Descriptive statistics of omissions in 52 DHS in sub-Saharan Africa.

	Negative		Positive		Mean	Minimum	Maximum
Categories	Signifi-cant	Not signifi-cant	Signifi-cant	Not signifi-cant			
All							
All women, all births	40	11	1	0	-0.09	-0.23	0.03
Education							
No education, all births	45	5	2	0	-0.06	-0.26	0.02
Some education, all births	25	21	5	1	-0.12	-0.30	0.08
Gender							
All women, male births	36	15	1	0	-0.10	-0.23	0.03
All women, female births	34	14	4	0	-0.09	-0.24	0.03
Survival status							
All women, surviving children	38	12	2	0	-0.09	-0.24	0.04
All women, deceased children	25	18	7	2	-0.15	-0.50	0.65

Table 3 also shows that relative omissions are on average twice as large among uneducated women as among educated women. The omission parameter is significantly negative among women with no education in 45 of the 52 surveys (as opposed to 25 surveys among the educated). This clearly suggests that the level of education of the population is a significant factor in explaining omissions. One of the consequences of this result is that fertility

[16] For analyzing omissions of deceased children, births are recoded as zero. We consider surviving in the first 6 months of life.

[17] The other parameters (displacements of births before or after the cut-off year) vary in a less systematic way and are less likely to be statistically significant than omissions, but are overall in the expected direction. The DB parameters tend to be significantly positive (in 18 surveys, indicating overestimation of fertility in the year before the cut-off year) and the DA parameters are more often negative (14 surveys), indicating underestimation of fertility during the cut-off year of the health module.

[18] The only positive case is the 2003 Kenya survey, suggesting that fertility may have been overestimated in the 2003 survey.

differentials by level of education are underestimated. Deceased children are also more likely to be omitted than surviving children. The omissions parameters are less often significant among deceased children, because of larger sampling errors – but the average level of omissions is greater among them (15%) than among surviving children (9%). This is expected, as it probably 'easier' to omit a deceased child, notably because the child was not listed in the household questionnaire. It is also in line with the idea that people (interviewers, respondents) may be embarrassed by questions about deceased children and may be more likely to omit deceased children (Sullivan, 2008). One of the consequences of this result is that child mortality is most probably underestimated in many surveys in sub-Saharan Africa. Finally, omissions do not vary by gender: about 9 to 10% of males and females births are omitted on average.

Are omissions important in explaining discrepancies between TFRs across surveys? A regression of the relative difference in TFR between surveys (dependant variable – see section 3) on omissions and displacements (independent variables) shows that omission is the only factor explaining differences between successive surveys. This can be explained by the fact that TFRs are computed for the three years before the survey, while the cut-off year of the health module is usually 5 to 6 years before the survey. In such a case, displacing births backward in time will have no effect on the TFR. In contrast, omissions are likely to occur regardless of the timing of births after the cut-off year.

In summary, these analyses show that omissions are widespread, and are a major issue in the measurement of fertility levels in sub-Saharan Africa (much more important than displacements of births). Given that omissions may vary from one survey to the next, or across subgroups, it also has implications on the measurement of fertility trends and fertility differentials.

6. Explaining omissions of births

Births can be omitted for a variety of reasons, either deliberately or by accident. Deliberate omissions means there must be some incentive for interviewers or respondents to omit these births. The most obvious incentive for an interviewer to omit births is to reduce his/her workload. Each birth that occurred in the reference period for the health module leads to a lengthy battery of questions. Omitting one or several birth thus leads to a gain of time. Another incentive is to avoid embarrassing questions. For instance, interviewers may feel embarrassed by question about deceased children, and may be more likely to omit deceased children in order to avoid these questions. Deliberate omissions of births may also be due to respondents, who may be reluctant to mention recently deceased children. One can also imagine that - like the interviewers - some respondents may omit recent births in order to gain time[19]. Finally, births may also be omitted by accident[20], but there is little reason to believe that recent births would be more likely to be omitted than births in the past.

There may also be a cost to omitting births. For the interviewer, the most obvious cost is to be sanctioned. This is more likely if the supervision of interviewers is strong. However, detecting systematic omissions during fieldwork is probably difficult, and the cost of omitting births for interviewers is probably low. The cost of omitting births for the respondents is also likely to be small.

Hypotheses and independent variables

Determinants of omissions are identified through macro-level analyses, by regressing relative omissions in each of the 52 surveys on a series of independent variables. These variables and the corresponding hypotheses are presented below. Summary statistics for these variables are presented in Table 4.

Length and complexity of questionnaires. The number of modules and the number of questions in the individual female questionnaires in DHS have increased significantly since the late 1980s, with a corresponding increase in the average duration of interviews. Our hypothesis is that interviewers may be more willing to omit some births and shorten the interview if the questionnaire is long and complex. Several factors could explain that. First, there may be more pressure on the interviewers to work quickly when the questionnaire is long than when the questionnaire is short. Another reason is that a long interview may be tiring both for the interviewer and the respondent, giving an incentive for the interviewer to shorten the interview. The length of questionnaire is measured by the number of modules (in addition to the standard modules) in the female survey. It both measures the length and the complexity of the questionnaire[21].

---

[19] Unless the respondents know the questionnaire, it is unlikely that they will be more likely to omit recent births in order to gain time. It is possible, however, that the information about the questionnaire circulates among women in a household or community, and that women who know the questionnaire tend to underreport recent births. This means that omissions should be lower among women interviewed first in the household (or community) compared to women interviewed later.

[20] For instance, respondents may not understand that a child who died quickly after his birth should be mentioned in the birth history.

[21] Two other variables measuring the length of the questionnaire give very similar results (results not shown). The first one measures the number of questions in the questionnaire, approximated by the number of non-empty variables in the recode data file. The second variable expresses the length of the questionnaire as a standardized average duration of interview. The number of modules is highly correlated with the number of questions (r=0.8), and with the average duration of the interview (r=0.7).

Table 4: Summary statistics of independent variables

	Mean	Std. dev.	Min	Max	N
Length and complexity of questionnaire					
Number of additional modules in the female survey	5.4	4.1	0	15	52
Length of health module					
Time cost of a birth in the reference period (minutes)	8.7	1.7	5.3	15.1	52
Reference period of the health module					
Short reference period (<42 months)	0.10	0.3	0	1	52
Size of the survey					
Number of interviewers	63.4	33.7	29	190	52
Number of interviews per interviewer	121.7	34.8	57.4	197.5	52
Education					
Percentage of population with some education	57.3	26.0	10.7	93.3	52

Length of the health module. A straightforward hypothesis to explain omissions is that omitting births is a strategy used by interviewers to avoid the lengthy health modules (on children) in the DHS (usually restricted to births in the last 5 years). According to this hypothesis, omissions of births should be more frequent in surveys in which the health module is longer, as the gain for the interviewer would be greater. The potential gain by omitted birth is measured by the 'time cost' of an additional birth in the reference period. The time cost per birth is measured for each survey by regressing the duration of interviews on the number of births for each interviewed women during the reference period. The 'time cost' of an additional birth is the increase in the duration of interview associated with an additional birth in the reference period. On average, a birth in the reference period lengthens the interview by 8.7 minutes, but this varies a lot across surveys (from 5.3 minutes to 15.1 minutes)[22].

***Reference period of the health module*.** While the health module is often restricted to births in the last 5 years, the reference period has varied significantly across surveys (from 38 months in the 1998 Cameroon DHS to more than 6 years in several surveys, like in Madagascar 2008). Our hypothesis is that a short reference period increases relative omissions (the dependant variable). This is based on the idea that some interviewers may be tempted to omit one birth, but are less likely to omit several births (notably because this would be too obvious). Omitting one birth in a short reference period will have a larger relative impact on fertility than omitting one birth in a longer reference period. Another possible reason is that incentives for omitting a birth may be greater with a short reference periods than with a long reference period. This is related to the fact that there are usually

[22] Surviving births and deceased births were also distinguished, as the number of questions varies depending on the survival status. The larger number of questions for surviving children translates into a longer average duration. These variables lead to very similar results, and are not presented here.

more questions about the latest birth than about previous births occurring during the reference period[23]. Omitting the most recent birth with a long reference period means that another birth will 'replace' the omitted most recent birth – and the gain will not be that large; In contrast, omitting the most recent birth with a short reference period (e.g. 3 years) means that probably no births will be included in the health module, leading to a larger gain for the interviewer. The length of the reference period is measured by a dummy variable distinguishing short reference periods (lesser or equal to 42 months) and long reference periods[24]. Among the 52 surveys, 6 (around 10%) had a short reference period.

***Size of the survey*.** Although larger sample sizes lead to smaller sampling errors, they may also increase non-sampling errors (including omissions) in several ways. A large sample size means that either the number of interviewers has to be increased, or the average number of interviews per interviewer must be larger. The second option means the surveys lasts longer; interviewers may be less motivated at the end of the fieldwork, and may be more likely to omit births. The other option (increasing the number of interviewers[25]) may also have detrimental effects on data quality. One reason is that is more difficult to recruit, train and supervise a large number of interviewers, and as a result to maintain high quality standards. Two variables are included in the models: the number of interviewers, and the average number of interviews per interviewer.

Level of education of the population. The level of education in the population may influence omissions in several ways. First, knowledge of dates is often less reliable in less educated populations, and the understanding of questions and definitions may be less developed among lowly educated women. This could leave more room for interviewers to omit some birth. Another possible influence of education on data quality is through the level of education of interviewers. In countries with low level of education, the recruitment of qualified interviewers may be more difficult than in countries with a larger pool of educated people. Education is measured by the percentage of the population (15 and over) with at least some primary education.

---

[23] On average, data collection in the health modules lasts about 8 minutes longer for the latest birth than for another birth in the reference period.

[24] The limit of 42 months corresponds grossly to a birth interval that would not seem exceptionally long. In other words, we consider that in reference periods shorter than 42 months, it is frequent that a woman has no birth, even without omissions. This means that birth omissions will be difficult to detect by supervisors, and will lead to a significant gain of time by interviewers.

[25] In some situations, a large number of interviewers may be selected for reasons not related to the sample size. For instance, because of problems of transportation, teams of interviewers may be recruited and trained in different regions.

Other factors could of course influence omissions. For instance, the quality of the training of interviewers is potentially an important factor, as is the quality of the supervision of the survey. Unfortunately, the information is not readily available and cannot be taken into account.

Results

Seven models are fitted to explain the overall level of omissions, and omissions by level of education, gender and survival status (Table 5). Standardized coefficients are presented to facilitate comparisons across variables. Negative coefficients indicate that a positive increase in the variable leads to more omissions.

Table 5: OLS regression of relative omissions of births in 52 DHS in Sub-Saharan Africa (by education), standardized coefficients.

Explanatory variables	All women & births	Education		Gender		Survival status	
		No	Some	Girls	Boys	Deceased	Surviving
Length of questionnaire							
Number of modules	-0.37***	-0.35***	-0.36***	-0.35***	-0.33***	0.13	-0.43***
Length of health module							
Time cost of birth in reference period	-0.05	-0.10	-0.05	-0.06	-0.04	0.16	-0.09
Reference period							
Reference period <= 42 months	-0.48***	-0.57***	-0.33***	-0.53***	-0.36***	-0.25**	-0.46***
Size of the survey							
Number of interviewers	-0.21+	-0.03	-0.29*	-0.09	-0.30**	-0.26+	-0.17+
Interviews per interviewer	-0.13	-0.18	-0.09	-0.07	-0.18+	-0.19	-0.09
Education							
% with some education	0.50***	0.13	0.33**	0.48***	0.44***	0.32**	0.46***
Adjusted R^2	0.53	0.43	0.41	0.48	0.43	0.25	0.50
N	52	52	52	52	52	52	52
Significance : ***: p<0.01; **: p<0.05; *: p<0.10; ; +: p<0.20							

Three variables are strongly correlated to omissions in the first model (all women and all births), and account for more the 50% of the total variance. First, a longer and more complex questionnaire leads to significantly more omissions. This result indicates that reducing the number of questions (and modules) in DHS may be a sensible strategy to improve data quality. Interestingly, this was already mentioned by Arnold (1990) as way to improve data quality, but the past twenty years have instead been characterized by increasingly long and complex of questionnaires. A second clear result is that a short reference period for the health module leads to larger relative omissions. This expected result is also very robust and clearly suggests that using a long reference period for the health module is a sensible approach to reduce omissions of births. DHS has favored long reference periods in most surveys, but short reference periods have been used in several cases (Cameroon 1998; Kenya 1998; Nigeria 1999; Mozambique 1997…), with detrimental effects on the quality of birth histories. Finally, as expected, omissions are much lower in countries with a higher percentage of people with some education, reflecting the fact the fact that omissions are larger among uneducated women.

Surprisingly, the length of the health module (measured by the time cost of a birth in the reference period) is not significantly correlated to omissions[26]. This does not mean that the health module is not responsible for omissions. Actually, the fact that fertility drops sharply after the cut-off year of the health module indicates that the existence of omissions is strongly related to the existence of the health module. However, variations in length of the health module do not explain variations in omissions. Omissions depend much more on the size of the questionnaire rather than the size of the health module in itself. The number of interviews per interviewer is not related to data quality. Finally, the number of interviewers is also related to omissions in the expected direction (a large number of interviewers leading to more omissions), but the coefficients are not significant in the all women model (but the p-value close to 0.1).

The conclusions are broadly similar in the 6 models for subgroups of births or women. There are, however, some differences. First, as expected, the percentage of women with primary education does not influence omissions among uneducated women. This indicates that impact of education on omissions in the overall population is explained by more frequent omissions among uneducated females. An impact of education on omissions is found among women with some education. This probably results from a composition effect (the percentage of women with secondary education may vary within the group of women with some education).

Secondly, the impact of a short reference period varies (significantly) between subgroups. Female births are more likely to be omitted in a short reference period than are male births. This may be explained by the fact that some questions (e.g. on female genital cutting) only concern young girls, and omitting a female birth may allow avoiding such questions. The short reference period also has a stronger effect among uneducated women, and among surviving children. A possible explanation is that interviews may be longer among uneducated women and for surviving children; omissions would lead to greater gains in these cases.

Third, the number of interviewers also has different effects on omissions depending on the models. Hiring a large number of interviewers has a detrimental effect on omissions of births

[26] Other indicators (such as the number of non empty variables in the health modules) lead to similar findings.

among educated women, male births and (to a lesser extent) deceased children. No satisfactory explanation was yet found for this finding. It means, however, that having a large number of interviewers may have a negative effect on data quality. Finally, the size of the questionnaire is not related to omissions of deceased children. This suggests that omissions of deceased children is not necessarily related to a gain of time, but rather reflect embarrassment.

7. Conclusion

This paper has shown that fertility estimates from DHS in sub-Saharan Africa are affected by serious data quality problems. Comparisons of fertility estimates from consecutive surveys indicate that recent fertility is virtually always underestimated. Although the idea that births histories are affected by data quality problems is not new, this paper shows that it is systematic and not negligible, and that omissions are the major source of underestimation. Omissions lead to underestimation fertility by about 10% on average – by comparison, confidence intervals for TFRs in DHS in sub-Saharan Africa usually also represent on average 10% of the TFR. Our results further indicate that omissions vary by education and survival status.

The determinants of omissions indicate that the design of the questionnaire (length of the reference period) can have a major impact on omissions. They also suggest that bigger surveys (longer questionnaire and large number of interviewers) tend to be lower quality surveys. In view of these results, we believe the balance between the quantity and quality of information in DHS should be reevaluated. Decreasing the size of the questionnaire would probably allow interviewers to concentrate on a smaller number of questions, without being incited to omit births. A smaller questionnaire would also allow hiring fewer interviewers, and may also improve data quality in this way.

Further analyses at the individual-level would provide additional evidence of factors influencing omissions. For instance it would be possible to evaluate the impact of the number of modules on omissions in the surveys in which an additional module was randomly attributed to a sub-sample of women. This random assignment means systematic differences in fertility could be attributed to differences in length of the questionnaire. Many other potential factors affecting omissions could be evaluated with individual data (timing of the interview, order of survey in the household…), and provide guidelines for improving data collection of birth histories.

8. References

Andersen, R. (2009). “Nonparametric Methods for Modeling Nonlinearity in Regression Analysis”, *Annual Review of Sociology*, vol. 35, pp. 67-85.

Arnold, F. (1990). “Assessment of the quality of birth history data in the Demographic and Health Surveys”, in IRD, *An Assessment of DHS-I Data quality*, DHS Methodological Report, n°1.

Blacker, J. (1994). “Some thoughts on the evidence of fertility decline in Eastern and Southern Africa; review of National Research Council, Population Dynamics of Sub-Saharan Africa”, *Population and Development Review*, vol. 20, n°1, pp. 200–205.

Choi, Y. and M. Sudhinaraset (2010). “Differential reporting errors in birth histories by socioeconomic group”, Abstract of Poster presented at the 2010 PAA Meeting, Dallas.

Cleland, J., Onuoha, N. and Timæus, I. (1994) “Fertility change in sub-Saharan Africa: a review of the evidence”, *in The Onset of Fertility Transition in Sub-Saharan Africa*, T. Locoh and V. Hertich (eds). Ordina Editions, Liège, pp. 1-20.

IGCME (2009). “Child mortality estimation in countries with high HIV/AIDS prevalence. Reference year of estimates: 2008”, *Inter-agency Group for Child Mortality Estimation (IGME)*, 2009, WHO, Geneva.

Kreyenfeld, M., A. Hornung, K. Kubisch, I. Jaschinski (2010). “Fertility and union histories from German GGS data: Some critical refl ections”, *MPIDR Working Paper*, n°33, MPIDR, Rostock.

Merli, M.G. and A. Raftery (2000). “Are Births Underreported in Rural China? Manipulation of Statistical Records in Response to China's population policies”, *Demography*, vol. 37, n° 1, pp. 109-126.

Narasimhan R., Retherford R. D., Mishra V., Arnold F. and Roy T. (1997), “Measuring the Speed of India’s Fertility Decline”, *National Family Health Survey Bulletin,* 6, IIPS & East West Center Program on Population, Mumbai & Honolulu

National Population Commission – *Nigeria (2000), Nigeria Demographic and Health Survey 1999*, National Population Commission, Abuja.

Ortega, J.A. (2009). “The fertility transition in sub-Saharan Africa: Early transition rise and (slow) decline”, *Paper presented at the 2009 PAA*, Detroit.

Pullum, T. (2006). “An assessment of age and date reporting in the DHS Surveys, 1985-2003”, *DHS Methodological Reports*, n°5, Macro International, Calverton, 93 p.

Retherford, R., G. Mirza, M. Irfan, and I. Alam, (1987). “Fertility trends in Pakistan – the decline that wasn't”, *Asian and Pacific Population Forum*, vol. 1, n°2, pp. 1-10.

Schoumaker B. (2004). “A person-period approach to analyzing birth histories”, *Population-E*, vol. 59, n°5, p. 689-702.

Schoumaker B. (2008). “Stalls and reversals in fertility transitions in Sub-Saharan Africa. Real or Spurious ?”, *IUSSP Seminar on Human Fertility in Africa*, Cape Coast, Ghana, September 2008.

Schoumaker B. (2010). “Reconstructing Fertility Trends in Sub-Saharan Africa by Combining Multiple Surveys Affected by Data Quality Problems”, *Paper presented at the PAA annual meeting,* Dallas, April 2010, 34 p.

Sullivan, J., Bicego G., Rutstein S. (1990). “Assessment of Quality of Data Used for the Direct Estimation of Infant and Child Mortality in the Demographic and Health Surveys”. In IRD, *An Assessment of DHS-I Data Quality*. Institute for Resource Development/Macro Systems, DHS Methodological Reports, 1. Columbia, Maryland.

Sullivan J. (2008). *An Assessment of the Credibility of Child Mortality Declines Estimated from DHS Mortality Rates*, A report submitted to UNICEF, Working Draft; Revision 1, 10/29/08.

Table annex 1. TFRs (first and second surveys) in the last three years and estimated omissions in the 52 surveys

Country	Year	TFR1 15-44 (first survey)	TFR 15-44 (second survey)	Relative omissions
Benin	1996	5.83	6.79	-0.14
Benin	2001	5.46	6.25	-0.16
Burkina Faso	1993	6.25	8.33	-0.10
Burkina Faso	1998	6.28	7.10	-0.08
Cameroon	1991	5.64	6.58	-0.10
Cameroon	1998	4.69	5.98	-0.16
Chad	1996	6.26	7.96	-0.12
Ethiopia	2000	5.45	7.19	-0.23
Ghana	1988	6.13	6.12	0.03
Ghana	1993	5.00	5.02	-0.02
Ghana	1998	4.35	4.77	-0.10
Ghana	2003	4.28	4.52	-0.04
Guinea	1999	5.36	7.60	-0.16
Kenya	1988	6.51	6.79	-0.05
Kenya	1993	5.15	5.51	-0.05
Kenya	1998	4.59	5.23	-0.12
Kenya	2003	4.78	5.38	-0.04
Madagascar	1992	5.97	6.77	-0.01
Madagascar	1997	5.79	6.27	-0.01
Madagascar	2003	5.02	5.60	-0.15
Malawi	1992	6.41	6.94	-0.09
Malawi	2000	6.14	5.94	-0.10
Mali	1987	6.83	8.36	-0.15
Mali	1996	6.50	8.34	-0.15
Mali	2001	6.53	8.23	-0.15
Mozambique	1997	4.91	6.58	-0.19
Namibia	1992	5.20	5.07	-0.02
Namibia	2000	3.96	4.27	-0.02
Niger	1992	6.78	8.29	-0.10
Niger	1998	6.95	9.40	-0.17
Nigeria	1991	5.66	5.98	-0.02
Nigeria	1999	4.60	5.86	-0.22
Nigeria	2003	5.48	6.40	-0.11
RCI	1994	5.09	5.76	-0.11
Rwanda	1992	6.00	7.03	-0.09
Rwanda	2000	5.62	6.20	-0.05

Table annex 1 (continued). TFRs (first and second surveys) in the last three years and estimated omissions in the 52 surveys

Country	Year	TFR1 15-44 (first survey)	TFR 15-44 (second survey)	Relative omissions
Senegal	1986	6.21	7.78	-0.09
Senegal	1992	5.80	6.50	-0.09
Senegal	1997	5.52	6.02	-0.05
Tanzania	1991	5.96	6.43	-0.06
Tanzania	1996	5.59	5.62	-0.09
Tanzania	1999	5.38	5.89	-0.09
Togo	1988	6.08	6.08	-0.05
Uganda	1995	6.66	7.46	-0.06
Uganda	1998	7.15	7.57	-0.03
Uganda	2000	6.61	7.82	-0.08
Zambia	1992	6.28	6.39	-0.07
Zambia	1996	5.92	6.67	-0.08
Zambia	2001	5.66	6.22	-0.10
Zimbabwe	1988	5.25	5.88	-0.07
Zimbabwe	1994	4.18	4.91	-0.15
Zimbabwe	1999	3.86	4.38	-0.08