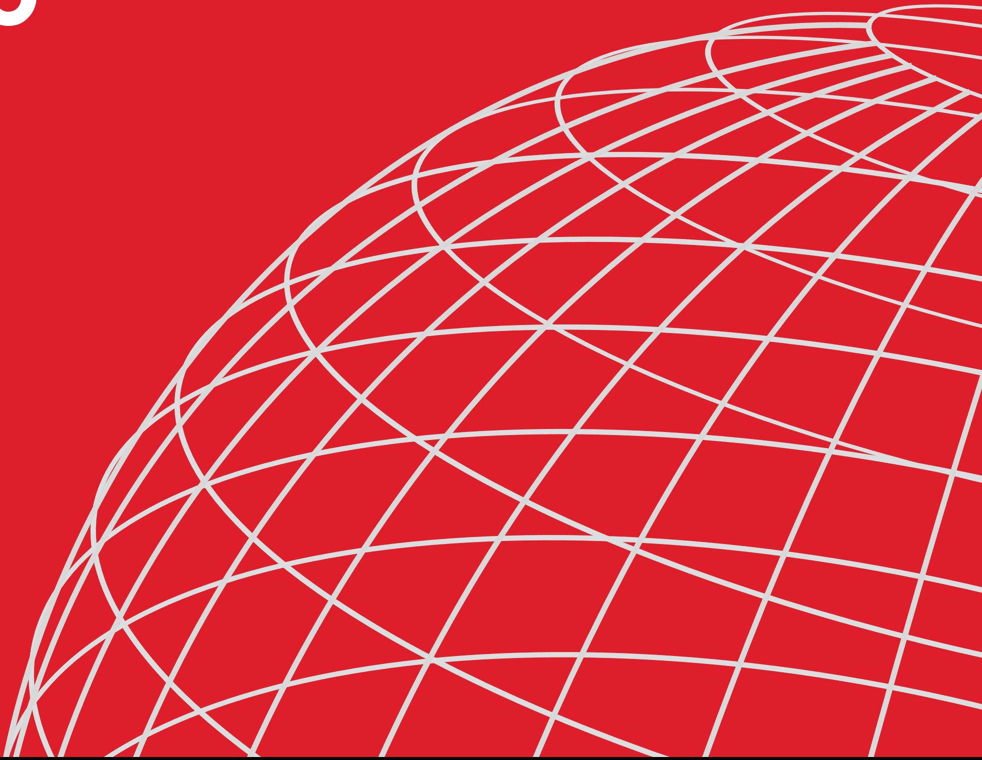
USAID
FROM THE AMERICAN PEOPLE

STRATEGIES TO ASSESS THE QUALITY OF DHS DATA

DHS METHODOLOGICAL REPORTS 26



September 2019

This publication was produced for review by the United States Agency for International Development. It was prepared by Thomas W. Pullum.

DHS Methodological Reports No. 26

Strategies to Assess the Quality of DHS Data

Thomas W. Pullum

The DHS Program
ICF
Rockville, Maryland, USA

September 2019

Corresponding author: Thomas W. Pullum, International Health and Development, ICF, 530 Gaither Road, Suite 500, Rockville, MD 20850, USA; phone: 301-407-6500; fax: 301-407-6501; email: tom.pullum@icf.com

Acknowledgments: The author thanks Stan Becker and Trevor Croft for very helpful comments on an earlier draft of this report.

Editor: Diane Stoy
Document Production: Joan Wardell

This study was conducted with support provided by the United States Agency for International Development (USAID) through The DHS Program (#720-OAA-18C-00083). The views expressed are those of the authors and do not necessarily reflect the views of USAID or the United States Government.

The DHS Program assists countries worldwide in the collection and use of data to monitor and evaluate population, health, and nutrition programs. Additional information about The DHS Program can be obtained from ICF, 530 Gaither Road, Suite 500, Rockville, MD 20850 USA; telephone: +1 301-572-0200, fax: +1 301-572-0999, email: info@DHSprogram.com, internet: www.DHSprogram.com.

Recommended citation:

Pullum, Thomas W. 2019. *Strategies to Assess the Quality of DHS Data*. DHS Methodological Reports No. 26. Rockville, Maryland, USA: ICF.

CONTENTS

FIGURES

PREFACE

The Demographic and Health Surveys (DHS) Program is one of the principal sources of international data on fertility, family planning, maternal and child health, nutrition, mortality, environmental health, HIV/AIDS, malaria, and provision of health services.

One of the objectives of The DHS Program is to continually assess and improve the methodology and procedures used to carry out national-level surveys as well as to offer additional tools for analysis. Improvements in methods used will enhance the accuracy and depth of information collected by The DHS Program and relied on by policymakers and program managers in low- and middle-income countries.

While data quality is a main topic of the DHS Methodological Reports series, the reports also examine issues of sampling, questionnaire comparability, survey procedures, and methodological approaches. The topics explored in this series are selected by The DHS Program in consultation with the U.S. Agency for International Development.

It is hoped that the DHS Methodological Reports will be useful to researchers, policymakers, and survey specialists, particularly those engaged in work in low- and middle-income countries, and will be used to enhance the quality and analysis of survey data.

Sunita Kishor
Director, The DHS Program

ABSTRACT

The Demographic and Health Surveys (DHS) Program strives to maintain the highest standards of data collection, processing, and analysis. This report is one of a series of DHS Methodological Reports on data quality. Earlier reports included descriptions of methods but focused on actual assessments of, for example, maternal mortality data or potential interview effects. This report focuses primarily on strategies and new methods, or significant modifications of methods that have appeared previously. The report includes many examples for illustrative purposes. The methods can be generalized to other substantive outcomes. For example, a chapter on the analysis of fertility that uses retrospective birth histories could be extended to under-5 mortality using the birth histories, or to adult and maternal mortality using the retrospective sibling histories.

The report has five main chapters. After the introductory chapter, Chapter 2 focuses on a type of displacement of birthdate, observed in many DHS surveys, that results from recording the year of birth as the year of interview minus years of age. The calculation is correct only if the respondent has already reached their birthday in the year of interview. Displacement of month of birth has been noted in the past but without an explanation of the mechanism behind it. Chapter 2 describes a simple indicator to measure the resulting bias, as well as other indicators based on the stated month of birth, and for children, the day of birth.

Chapter 3 focuses on the quality of the birth histories. The main method is a comparison of two successive surveys and their estimates of fertility rates for the 5 calendar years before the first survey. Single-year rates, as well as 5-year rates, are compared with statistical models. Chapter 4 uses statistical models to describe variations in data quality indicators according to characteristics such as place of residence, household wealth, and level of education. Quality-related outcomes, such as nonresponse or age heaping, may potentially vary for reasons that lie beyond the implementation of fieldwork. It could be argued that interviewer effects, for example, should be adjusted for the characteristics of the individuals being interviewed.

Chapter 5 discusses the interview process and, in particular, the duration of the household interview, as related to its position within the duration of fieldwork, within the duration of fieldwork in a specific cluster or time of day, and how the duration of the interview relates to the number of household members and the number of items in the questionnaire. In the same way as age heaping, very short interviews may suggest substantively serious data quality problems. Chapter 6 provides examples of how it is possible to focus on interviewers, clusters, and days, in any combination, in which there was an irregularity that was both large in magnitude and statistically significant. It is possible to simulate or track the fieldwork in a variety of ways by using information from the data files.

The methods described here, as well as those that appeared earlier, will be used in the future to prepare data quality profiles of all DHS surveys, and to better monitor long-term trends in data quality and identify potential problems in new surveys. Some methods can also be adapted to better monitor data quality in real time during fieldwork.

Key words: data quality

ACRONYMS AND ABBREVIATIONS

AIDS	acquired immunodeficiency syndrome
CAPI	computer-assisted personal interviewing
cdc	century day code
DOB	date of birth
DHS	Demographic and Health Surveys (The DHS Program)
DQ	data quality
HAZ	height-for-age
HIV	human immunodeficiency virus
KIR	key indicators report
MICS	Multiple Indicator Cluster Survey
MOB	month of birth
MMR	maternal mortality ratio
MR	methodological report
NR	nonresponse
OD	other document
OLS	ordinary least squares
TFR	total fertility rate
WAZ	weight-for-age
WHZ	weight-for-height
WP	working paper
YOB	year of birth

1 INTRODUCTION

The Demographic and Health Surveys (DHS) Program continuously strives to maintain and improve the quality of DHS data. This report will describe and illustrate a set of procedures, grounded in demographic and statistical methods, that measure the quality of data after they have been collected. Together with procedures described in other reports on data quality, they form a framework that can be applied to every survey, using the standard recode files, to position that survey relative to other surveys in the same country or in other countries.

These methods complement other monitoring approaches that are applied during data collection, rather than afterwards, in order to identify potential misinterpretations and patterns of errors by fieldworkers. The methods can lead to improvements in questionnaire design and the training and supervision of fieldworkers.

Our goal is not to summarize the quality of a survey with a single number, because there are theoretical and empirical reasons to believe that quality is multidimensional. At the same time, it is not helpful to produce a plethora of indicators that are difficult to interpret and translate into improvements during data collection. The strategies are designed to systematically address potential measurement errors in important components and outcomes of DHS surveys.

Five themes will be considered in successive chapters:

Chapter 2. Age and birthdate
Chapter 3. Fertility
Chapter 4. Variations related to characteristics of the respondent
Chapter 5. Variations related to duration of the interview
Chapter 6. Variations during fieldwork

Each chapter will be relatively self-contained with its own methodology and illustrative examples. A final chapter will offer conclusions and describe further extensions.

With only a few exceptions, the indicators of data quality are calculated at the level of the individual and are then analyzed with statistical methods that can produce standard measures of dispersion, confidence intervals, and statistical tests. Individual-level indicators can be correlated with one another and included in statistical methods such as logit regression and principal components analysis. Such indicators can also be used to construct summary indices familiar to demographers such as age ratios and measures of age heaping or digit preference. The transition of data quality indicators to a statistical format is analogous to the ongoing statistical reformulation of the demographic and health indicators that are produced for the main reports and STATcompiler.

One of the analytical possibilities facilitated by individual-level versions of data quality indicators is the opportunity for multivariate analysis, using as covariates the characteristics of the respondents such as place of residence, region, level of education, and wealth quintile. Multivariate analysis could be applied to almost any topic in the report.

The general approach to a statistically based investigation of data quality has the following components for a given indicator and survey:

1. Calculate an expected level of the indicator, based on documented relationships, similar surveys from other countries, an earlier survey from the same country, or data within the given survey.
2. Calculate the deviation of the indicator from the expected level, if in a worse direction, for respondents, interviewers, or teams.
3. Construct a statistic to test whether this deviation is statistically significant, taking into account the number of cases on which it is based and a null hypothesis that the deviations are random.
4. Specify a tolerance level, or threshold, for the indicator, which is somewhat arbitrary but should be below the level at which substantive inferences would be jeopardized.
5. Flag cases for which the observed level exceeds the expected level by a highly significant amount and exceeds the tolerance or threshold level.

This approach is generally followed in this report. For example, one might wish to identify interviewers whose reports of age are poor. The prevalence of age heaping is an indication of the quality of age reporting. The measure of heaping could be the percentage of ages that end in final digits 0 or 5. We would expect 20% of ages to end with 0 or 5, because these are two of the ten digits 0 through 9. A tolerance level could be set at 30%, for example. We then could flag interviewers whose observed percentage exceeds the threshold (30%) *and* exceeds the expected value (20%) with a p-value of .001.

Because there are many interviewers and each interviewer generally conducts many interviews, a .001 criterion would be preferable to .05 or .01. We do not test whether the prevalence significantly exceeds the tolerance threshold, but whether it exceeds the expected value. The preferred analysis for a binary indicator such as this would use logit regression, which can also identify *avoidance* of final digits 0 and 5, which can occur if interviewers have been overtrained to avoid heaping on these digits.

These components of the general strategy will be reviewed in different chapters of this report, although occasionally with different terms or a different sequence.

Some clarification of the third step, statistical testing, is desirable. The null hypothesis is simply that deviations from an expected value are random. These tests do not have the same kind of interpretation as, for example, a test of whether an indicator such as the maternal mortality ratio (MMR) has declined between two surveys or whether an intervention to reduce neonatal mortality has been effective. For these more substantive examples of hypothesis testing, the costs of an incorrect inference can be considerable.

Decisions based on indicators of data quality are limited to (a) releasing the data files and producing analyses based on the data, without casting doubts on data quality, or (b) releasing the data files and producing analyses, but going beyond the normal caveats and advising caution, or (c) suppressing some of the data or even the entire survey. Very few surveys have been restricted because of data quality concerns. The most recent example was a 2017 survey in Niger. A special report (OD73) provides a justification of the decision to suppress that survey. Several surveys have had partial suppression of data, such as anthropometry data, or data from a specific region. The reports of a few surveys have included strongly worded cautions about some indicators such as those related to under-5 mortality or fertility.

In general, two types of potential errors can occur when making a decision about releasing or withholding data when there are data quality concerns. In the present context, a Type I error is made if the data are suppressed or questioned when in fact they are satisfactory. For example, the maternal mortality data may indicate an implausibly high increase in maternal mortality, and as a result, the estimate is suppressed, but it is actually correct. When DHS encounters a potential Type I error, the usual action is to conduct a further investigation and add caveats, but not suppress the data—that is, to follow option (b) above rather than option (c). Users of the data, including DHS staff, can subsequently reanalyze and perhaps adjust the data. After a later survey, the estimate may be evaluated again. Option (c) is rarely taken.

Type II error occurs when the data are not suppressed or questioned, but in fact (at least from a strictly data quality perspective) they should be. A relatively lower level of Type I error necessarily implies a relatively higher level of Type II error. By keeping Type I errors low, DHS tends to release data that are questionable and to refrain from highlighting concerns, which may mask the fact that there was considerable internal discussion before the release of the data or a report. DHS has begun attaching confidence intervals and more routinely describing the uncertainty inherent in all estimates. There has been an effort to sensitize users to the presence of both sampling and nonsampling errors.

The relative bias toward releasing data known to be flawed, but offering cautions and caveats, when appropriate, is arguably a better option than suppressing data entirely. Such a policy, combined with routine quality checks during and after data collection, is the pathway to achieving better data in future surveys. Lessons learned and discussed openly can lead to improvements. We suggest that the costs and other negative implications of suppressing questionable data are greater than the costs of making available data that may be questionable, as long as the data release is accompanied by a thorough assessment of quality.

The specific surveys used to illustrate the strategies in this report are drawn from standard DHS surveys conducted since 2000. Sometimes two surveys will be used for a data quality (DQ) indicator, one illustrating high levels and one illustrating low levels. The selection of specific surveys is not random, but it is not based on an intention to represent any survey as outstandingly poor or outstandingly good. The intent is simply to illustrate that the indicators take a range of values and can discriminate across surveys. Some empirical evidence of the distribution of the indicator and patterns of association, across many surveys, is included for some data quality indicators.

Many earlier reports, primarily methodological reports (MRs), but including one working paper (WP) and one other document (OD), have focused on specific components of data quality:

MR1 (1990):	An Assessment of DHS-I Data Quality
MR2 (1994):	An Assessment of the Quality of Health Data in DHS-I Surveys
MR5 (2006):	An Assessment of Age and Date Reporting in the DHS Surveys, 1985-2003
MR6 (2008):	An Assessment of the Quality of Data on Health and Nutrition in the DHS Surveys, 1993-2003
MR11 (2014):	Evidence of Omission and Displacement in DHS Birth Histories
MR12 (2014):	Quality and Consistency of DHS Fertility Estimates, 1990 to 2012
MR13 (2014):	An Assessment of DHS Maternal Mortality Data and Estimates
MR16 (2015):	An Assessment of the Quality of DHS Anthropometric Data

MR17 (2015):	Contraceptive Use and Perinatal Mortality in the DHS: An Assessment of the Quality and Consistency of Calendars and Histories
MR18 (2017):	Hemoglobin Data in DHS Surveys: Intrinsic Variation and Measurement Error
MR19 (2017):	An Assessment of the Quality and Consistency of Age and Date Reporting in DHS Surveys, 2000-2015
MR21 (2017):	Comparisons of DHS Estimates of Fertility and Mortality with Other Estimates
MR24 (2018):	The Effect of Interviewer Characteristics on Data Quality in DHS Surveys
MR25 (2018):	Consistency of Reporting of Terminated Pregnancies in DHS Calendars
OD73 (2018):	Data Quality Evaluation of the Niger 2017 Demographic and Health Survey
WP162 (2019):	Evaluation of Indicators to Monitor Quality of Anthropometry Data during Fieldwork

The earlier reports will be referred to simply as MR1, MR2, etc. The bibliography provides complete citations.

With the exception of MR21, all reports listed above were actual assessments of data quality. The present report describes strategies to assess data quality, and is not an assessment. This report includes some new methods, and extends some methods introduced in the earlier reports, with minimal repetition of the methodologies already presented in those reports. The many examples illustrate strategies, rather than reaching conclusions about the specific surveys in the examples. These strategies will be systematically applied to a large number of surveys in a future report, for which the present report provides the methodological background.

Some topics and types of data are not included in this report. Anthropometry and hemoglobin measurement are important, but were considered in MR16, MR18, and WP162. The quality of the maternal mortality estimates was studied in MR13, while the quality of calendar data was the focus of MR17 and MR25. These topics will be included in future assessments of data quality, which will use the indicators and methods developed in the earlier reports.

2 AGE AND BIRTHDATE

2.1 Introduction

Sex and age are arguably the most important demographic characteristics of living individuals and are crucial for the calculation of almost all rates and indicators derived from DHS data. This chapter focuses on the reporting of age and birthdate. Age in years is obviously determined by birthdate, combined with the date of interview, but the reason for distinguishing between the two will become clear later in this chapter. This introduction briefly reviews the importance of correct information on age and some previously established ways to assess the age data.

The DHS Program is an important source of indicators of child health, most of which are explicitly related to age and/or birthdate. These include scores for height-for-age (HAZ) to determine the prevalence of stunting; scores for weight-for-age (WAZ) to determine the prevalence of underweight; the median duration of breastfeeding, which as a current status measure is based on current age; and age eligibility thresholds for immunizations, hemoglobin, and malaria testing. Recent fertility rates and under-5 mortality rates depend on the correct specification of birthdate, while the criterion for the entire set of under-5 questions and measurements is that the child was born within the 5 years before the day of interview.

Age is also critical for adults because it is a criterion for eligibility for the surveys of women and men. Age-specific fertility rates require estimates of a woman's age at the time of each child's birth, which are derived from her birthdate and the child's birthdate.

In many countries with DHS surveys, most births are not registered, exact dates of birth are not recorded, and birthdays are not observed. Since ages and birthdates estimated in the surveys cannot be corroborated against other sources, it is particularly important that the procedures used to estimate them during data collection be well understood and that the quality of the reported ages and birthdates be checked for internal consistency. The quality of the age data is important in its own right, but it is generally assumed to be associated with the quality of measurement of all indicators, including more complex indicators such as rates that are harder to assess directly.

Indicators have been developed to assess three potential weaknesses in the reporting of age: incompleteness, heaping, and displacement. Aggregate-level and individual-level indicators of all three were provided in MR19 and MR24, and will be reviewed only briefly here.

Incompleteness refers to whether age, year of birth, month of birth, and day of birth (if required, as for young children) are given and are consistent at the date of interview. For an aggregate such as women, men, or children, the percentage of cases with a code for incompleteness serves as an indicator. For individuals, the binary indicator is coded 1 for incompleteness and 0 otherwise, except that if the relevant code for incompleteness is NA, the indicator is also NA.

Age heaping can be interpreted as unevenness in the final digit of age or, more specifically, as an excess (or deficit) at final digits 0 or 5. For an aggregate it can be assessed with Myers' Blended Index. (MR5 developed an individual-level version of that index.) Myers' Blended Index is essentially the same as the Index of Dissimilarity, which is half the sum of the absolute values of the differences between observed and expected percentages at each final digit. The only difference between Myers' Blended Index and the Index of Dissimilarity is in the "blending," which takes into account unevenness resulting from the pyramidal shape of the age distribution. Either index can be interpreted as the percentage of cases that would have to be moved from overreported digits to underreported digits in order to achieve a uniform distribution.

A simple individual-level indicator used in MR19 and MR24 is defined to be 1 if the final digit is 0 or 5, and 0 otherwise. Heaping is observed if the indicator is 1 for more than 20% of cases, or for more than some threshold that is higher than 20%. Avoidance of 0 and 5 is identified similarly. The individual-level measure can be included in multivariate analysis to identify covariates that are related to heaping.

For both adults and children, age data can show evidence of displacement. The most serious type of age displacement for women age 15-49 is across the boundaries of the interval, at the 15th and 50th birthdays. For example, many surveys show an obvious deficit of 15-year-olds, which is probably the result of interviewers tending to report a girl as age 14, rather than 15, in order to reduce their workload. It is believed that this kind of displacement occurs when there is genuine uncertainty about a girl's age, in which case there are implications for other ages, not just 14 and 15. Transfers from age 49 to age 50 are confounded with heaping at age 50. For men, the upper age of eligibility may be 54 or 59, depending on the survey.

For young children, there tends to be displacement across the boundary for the child health questions, which is usually set at January of the fifth calendar year before the first calendar year of fieldwork. In some surveys, there is evidence that some children are displaced into an earlier year of birth so that the health questions will not need to be asked. Although this displacement is often described as a transfer from (completed) age 4 to age 5, it is actually a transfer from one birth year to another, and is a type of transfer of birthdate.

It is impossible to determine whether specific individuals have been shifted from age 15 to age 14 or from age 49 to age 50, or whether children have been shifted into an earlier year of birth. We would expect approximately equal numbers of cases at both ages or at both birth years. For analyses of data quality, we can construct individual-level indicator variables and include them in statistical models to determine which types of respondents are more likely to experience heaping and transfers.

The binary indicator of displacements or transfers from age 15 to age 14 is defined to be 1 if age is 15, 0 if age is 14, and NA otherwise. In logit regressions, a tendency to displace downward, out of eligibility, is indicated by a coefficient that is negative or an odds ratio less than 1. The binary indicator of transfers from age 49 to age 50 is defined to be 1 if age is 49, 0 if age is 50, and NA otherwise. In logit regressions, a tendency to displace upward, out of eligibility, is indicated by a coefficient that is negative or an odds ratio that is less than 1.

The binary indicator of birth year transfers is defined to be 1 if the birth year is the calendar year just inside the interval of eligibility for the health questions, 0 if it is the previous calendar year, and NA otherwise. A logit regression coefficient or an odds ratio will be negative if there is evidence of backward transfers.

It is possible to reverse the "1" and "0" categories of indicators of displacement; the effect is only to change the sign of the coefficient. Aggregate-level indicators are built up from the individual-level indicators.

2.2 Stated Age and Implied Age

This section is relatively lengthy and includes considerable detail on a new method to identify another type of potential misreporting of age, again interpreting its relative frequency as an indicator of the quality of age data more generally. In addition to an aggregate-level analysis, we include individual-level indicators that can be used in multivariate analyses. This type of misreporting can be described as a potential systematic displacement of birthdates that takes place during fieldwork. The displacement has been observed earlier, by Agarwal et al. (2017) and by Larsen, Headey, and Masters (2019), but without an explanation of its origin.

The fieldwork in every DHS survey begins with the preparation of a roster or list that includes all persons in the sampled households. For each person, the roster includes the person's name (which is entered at the time of data collection but is removed in the final stage of data processing) and the following characteristics: relationship to the household head; sex; whether the person is a usual resident; whether the person stayed in the household the previous night; and age in completed years, hv105.[1] All of this information is provided by the household respondent, who is typically either the household head or the spouse of the household head, and a parent of all or most of the children in the household. The household interview includes other information, but eligibility for virtually all questions about individuals depends on the initial assessment of sex and years of age. We will refer to years of age initially given in the household survey as stated age.

For many household members, month and year of birth are collected a second time, after the conclusion of the household interview. Thus, during the interview of eligible women age 15-49, the woman is asked for the month and year of her own birth and of each child she has ever had. In most surveys, a majority of the surviving children under age 15, say, of the interviewed women, will be in the same household as the woman and will appear in both the household roster and the mother's birth history. For children born during the past 5 years, the mother is also asked for the day, month, and year of birth. Beginning with surveys conducted in 2016, women are asked to provide a day of birth for all births, not just those in the past 5 years. During the interview of eligible men (the age range for men is usually 15-49, 15-54, or 15-59), the man is asked for his month and year of birth.

The day, month, and year of birth for children under age 5 are required for the calculation of the HAZ and WAZ scores. The effect of potential misreporting of age on these scores is one motivation, although not the only one, for developing the indicators in this chapter.[2] Since about 2000, DHS has measured the height and weight of all children in the household, not just those children whose mother resides in the household. If the household contains any children whose mother has died or resides elsewhere, their day, month, and year of birth are provided by the household respondent, after the completion of the household roster.

[1] To be as specific as possible, we will refer to variable names as they appear in the data files. Some technical details will be included in footnotes. Age in the household roster is hv105 in the household data file.

[2] b1 and hc30 (month of birth), b2 and hc31 (year of birth), hw16 and hc16 (day of birth) are duplicates in the PR and KR files that always match. b3 and hc32 (cmc of birth), hw1 and hc1 (months of age, if living), and b8 (years of age, if living) are calculated during data processing. Variables that begin with b and hw only appear for children whose mother is in the household. Variables that begin with hc appear for all children.

For all persons for whom the month and year of birth are obtained, during data processing there is a calculation of the person's age at the time of the survey based on the difference between the month and year of interview and the stated month and year of birth. For children whose birthdates include day, day is taken into account. The result of this calculation, in completed years, is the implied age. Most uses of age in DHS tabulations give priority to implied age.

We emphasize that all appearances of age in the recode files, other than hv105, are the result of data processing calculations that use the date of interview and date of birth as entered by the interviewer.[3] The age implied by the difference between the date of interview and the date of birth does not necessarily match with stated age.

Suppose, for example, that the household respondent is the child's mother, which is often the case. The mother says that the child is age 3. Subsequently, the interviewer negotiates the day, month, and year of birth. Suppose that the interviewer enters a birthdate that would imply that the child is age 4. Regardless of the magnitude of a discrepancy between stated age and implied age, after the birthdate has been assigned, the stated age is ignored for nearly all the data processing and analysis.

2.3 A Framework for Displacement of Birthdate

The easiest way to describe the potential pattern of displacement is with a specific example of a hypothetical child. Suppose that the date of interview is April 20, 2013. In the household interview on that date, a child is reported to be age 3. The interviewer is subsequently required to provide a year, month, and day of birth, so that the child's age can be calculated more precisely.

A possible strategy to localize the birthdate—if the correct age is indeed 3—would be to ask whether the child has already had a birthday in 2013. If the answer is yes, then that birthday must have been the 3rd birthday, so that the child turned 3 sometime between January 1 and April 20, 2013, inclusive, and was therefore born in the interval January 1 to April 20, 2010. Here 2010 is the result of a simple subtraction: 2013-3=2010. If the child is 3, but has not yet had a birthday in 2013, the child will turn 4 later in 2013 and was therefore born in the interval April 21 to December 31, 2009. Here, 2009 is calculated as 2013-4=2009.

The month and day of the interview, within 2013, provide information about whether a 3-year-old child was born in 2009 or 2010. Let P be the proportion of the year of interview[4] that has elapsed up to and including the day of interview. In this example, April 20 is the 110th day of the year and P = 110/365 = 0.301. The probability that the child has already had his/her birthday in 2013, and therefore was born in 2010, is approximately 0.301. The probability that the child was born in 2009 is approximately 1-.301 =

[3] In surveys conducted prior to 2016, children's age in months, hc1, is calculated during data processing as the difference between the cmc of interview and the cmc of birth. That is, hc1=v008-b3. Implied age in years is not given in the PR file but is given in the KR and BR files as b8; it is calculated from age in months as the integer part of hc1/12. In surveys conducted since 2016, which include day of birth, age in days is given as hc1a, which is consistent with day of birth and day of interview, and hc1 is calculated as hc1=int[hc1a/(365.25/12)].

[4] The reference is to the year of each specific interview and not the year of the survey. The interpretation is the same if the survey is contained in one calendar year, is spread across two years, or is continuous.

0.699. The only reason for including the qualifier “approximately” is possible seasonality of births and variations in child survival.

This child’s stated birthdate can be classified within one of the following six time intervals, T, numbered chronologically from earlier date to later date:

T=1: Year is earlier than 2009: The birthdate is completely inconsistent with stated age.

T=2: Jan. 1 - April 20, 2009: The year could be consistent with stated age, but not for these days.

T=3: April 21 – Dec. 31, 2009: The combination of days and year is consistent with stated age.

T=4: Jan. 1 - April 20, 2010: The combination of days and year is consistent with stated age.

T=5: April 21 – Dec. 31, 2010: Year could be consistent with stated age, but not for these days.

T=6: Year is later than 2010: The birthdate is completely inconsistent with stated age.

Two of these intervals, 3 and 4, are consistent with the stated age of 3 on April 20, 2013. Intervals 2 and 5 are consistent with a possible year of birth, but not with the range of days. Intervals 1 and 6 are not consistent with either of the 2 possible calendar years for this child, which are 2009 and 2010. If the child is displaced to intervals 1 or 2, the calculated years of age for the displaced birthdate will be 4 or more; if displaced to intervals 5 or 6, the calculated age will be 2 or less.

This pattern of displacement will now be described in a more general way, for a person of any stated age A in the household survey and any date of interview, specified as year Y and day D (D=1, 2, …, 365). The two consecutive birth years that could be consistent with age A will be labelled Y1=Y-A-1 and Y2=Y-A. In the previous example, Y=2013, A=3, Y1=2009, and Y2=2010. Define D1 to be an *interval* of days, and specifically the interval from January 1 to day D, inclusive. Define D2 to be the remainder of the year. P=D/365 is proportion of the year that is in interval D1. In the previous example, D1 is the 110 days from January 1 to April 20, inclusive; D2 is the 255 days from April 21 to December 31, inclusive; and P = 110/365 = 0.301. The stated age A is compatible with the combinations (D2,Y1) and (D1,Y2). Birthdates can be placed within any of six possible time intervals T, numbered chronologically from earlier date to later date, as follows:

T=1: Year<Y1: Inconsistent.

T=2: (D1,Y1): Y1 could be consistent with stated age, but not in combination with D1.

T=3: (D2,Y1): The combination of days and year is consistent with stated age.

T=4: (D1,Y2): The combination of days and year is consistent with stated age.

T=5: (D2,Y2): Y2 could be consistent with stated age, but not in combination with D2.

T=6: Year>Y2: Inconsistent.

Time intervals 2 and 3 are the first and last parts of year Y1; intervals 4 and 5 are the first and last parts of year Y2. Only intervals 3 and 4 are consistent with the stated age. Apart from possible seasonality of births, we would expect the odds of interval 4 versus interval 3 to be approximately the ratio of the days in D1 to the days in D2, i.e., P/(1-P). We propose a set of comparisons among the six intervals, and expectations about the relative frequencies of intervals 3 and 4 in the observed data. We will suggest scenarios under which the relative frequencies may differ from the expectations.

Sometimes, what may first appear to be an inconsistency is in fact the correction of an earlier error, and an improvement in data quality. For example, the household respondent could be someone other than the child's mother, and he or she misstates the child's years of age. When the mother is consulted, she may make a correction and provide a birthdate that is consistent with the corrected age. For adult women and men in the household, the household respondent may give an incorrect age that is corrected later during the individual interview with that adult. Indeed, it is expected that the birthdate, and the implied age, will be more accurate than stated age.

We suggest that there may be a tendency for interviewers not to ignore stated age, but to try to identify a birthdate that is consistent with stated age. They tend to do this correctly and to find a birthdate in intervals 3 or 4, but sometimes they miscalculate. The result may lie outside of intervals 3 and 4, or may show a preference for interval 4 over interval 3. We suggest that inconsistencies between implied age and stated age, whether they result from a correction or a miscalculation, tend to have a negative effect on the overall quality of the age data.

2.4 Indicators and Methods

Year of Birth Indicator YOB1, "shift_out". If the initial stated age is correct, all cases should be in intervals 3 and 4. If another interval occurs, there are two plausible explanations. First, cases in intervals 1, 2, 5, or 6 may result from the correction of initially erroneous statements of age. Second, the stated age may be correct but the birthdate is calculated incorrectly because of an arithmetic error. The first indicator of year displacement is the percentage of cases that are reported in intervals 1, 2, 5, or 6.[5] An arbitrary threshold for the seriousness of this indicator is 5%.

Year of Birth Indicator YOB2, "inner_up". Of the compatible classifications into intervals 3 and 4, we would expect with perfect data that the allocation between those two intervals will be related to the date of the interview, within the calendar year of data collection. For children interviewed a fraction P into the year, we expect that the probability of interval 3 is 1-P and the probability of interval 4 is P. If there is a deviation from this expectation, it is expected that birthdates will disproportionately tend to be placed in interval 4 rather than interval 3, simply because the year in interval 4 is given by the current calendar year minus age in years, which is a simple and intuitive calculation. The indicator is the log odds of interval 4 relative to interval 3, adjusted for the expected odds P/(1-P). A threshold for this indicator is +/-log(1.25) or +/-0.223. This threshold is selected after analyzing many surveys and inferring that the maximum displacement that can be attributed to genuine seasonality will increase or decrease the log odds of category 4 versus category 3 by about 25%. This threshold is described below as a tolerance.

[5] A p-value for YOB1 can be obtained with a logit regression with no covariates. The p-value is always very small. A confidence interval can also be obtained but for present purposes is not helpful.

Year of Birth Indicator YOB3, "inner_up_pvalue". We include the p-value of the test statistic for a null hypothesis that the balance between categories 3 and 4 is completely consistent with the expectation based on day of interview. The smaller the p-value, the less likely that the observed imbalance is due to an insufficient number of observations. This is a two-tailed p-value. The threshold for inferring displacement is p<.001, a very high standard.

YOB1 is calculated as the mean (multiplied by 100) of an individual-level binary variable called yob1.

YOB2 is the coefficient of a logit regression of an individual-level binary variable called yob2 (with no covariates and an offset). YOB3 is the p-value for that coefficient. The binary variables yob1 and yob2 are defined as:

yob1=1 if T=1, 2, 5, or 6; yob1=0 if T=3 or 4. YOB1 is the mean of yob1, multiplied by 100.

yob2=1 if T=4; yob2=0 if T=3; otherwise yob2=missing. YOB2 is the coefficient from a logit regression.

The logit regression using yob2 has no covariates and includes an offset, log[P/(1-P)]. YOB2, the coefficient from the logit regression, can be interpreted as the log of the odds of T=4 versus T=3, adjusted for the date of the survey within the calendar year. If the coefficient is greater than 0 (the odds are greater than 1), then there has been net displacement from T=3 to T=4. This is the hypothesized direction of displacement, based on an attraction toward year Y-A. However, because of seasonality, overtraining, or some other factor, there may be displacement in the opposite direction, which is indicated by a fitted log odds that is negative. Thus, deviations in either direction are important. YOB3 is the p-value for this logit regression.

Displacement from intervals 3 and 4 into intervals 1, 2, 5, or 6, or displacement from interval 3 to interval 4, will affect the reporting of the month of birth. For example, a shift from 3 to 4 will require reporting a month and day that are earlier in the year than the date of interview, when the true month and day are actually later in the year. This kind of displacement may be incorrectly interpreted as seasonality of births. The term seasonality, with respect to births, refers to a potential tendency for the distribution of birthdates to differ from a uniform distribution within a calendar year. Seasonality of births is a real phenomenon. Some variation across calendar months is to be expected, systematically across many years of births, or possibly specific to certain calendar years. Some unevenness results just from variation in the number of days per month. The challenge is to distinguish between genuine seasonality and spurious seasonality induced by displacement. The following indicators of displacement of month and day of birth are not affected by genuine seasonality, but can help identify spurious seasonality.

Month of Birth Indicator MOB1, "mob_moi_ind_pvalue". We test whether the calculated month of birth is independent of the month of interview. If month of birth is reported correctly, it should have no statistical relationship with the month of interview. The two variables are cross-tabulated and a chi-square test of independence is calculated. We do not report the chi-square statistic, but only the p-value for the chi-square statistic. This is the only indicator directly related to month of birth. It is an aggregate-level indicator with a threshold value of .001.

Displacement into interval 4 will induce a departure from independence between month of birth and month of interview, because the displaced births will have a birthdate that is earlier in the year than the date of interview. For that reason, there is a correspondence between MOB1 and YOB3, in which usually both will be non-significant or both will be highly significant.[6]

Day of Birth Indicator DOB1, "dob_uniform". We assume that, within a month, the true day of birth is uniformly distributed. The data often show preference for days 10 and 20, as multiples of 10, for example, and sometimes preferences for other multiples of 5, or for even-numbered days. Another pattern sometimes observed appears to be a preference for days in the first half of the month or earlier in the month than the day of interview. DOB1 describes variation across days 1-30, ignoring year and month.[7] It is an application of the index of dissimilarity, in which the distribution of days is compared with a uniform distribution. It is interpreted as the percentage of births in overreported days that would have to be shifted to underreported days to achieve a uniform distribution. An arbitrary threshold for this aggregate-level indicator is 10%.

Day of Birth Indicator DOB2, "dob_uniform_pvalue". A chi-square test of the null hypothesis that the days are equally likely (essentially a test of the null hypothesis that DOB1=0 in the population) is carried out with poisson regression. We do not actually provide the chi-square statistic but only its p-value. It is an aggregate-level indicator and has a threshold value of .001. DOB1 and DOB2 are restricted to cases for which day of birth is coded in the data files.

The day of birth indicators can generally be calculated only for children age 0-4. Beginning in 2016, DHS surveys include day of birth in the birth histories for all children. Thus, for recent surveys these indicators can also be calculated for children age 5-14 who are in both the household survey and the mother's birth history.

We do not take into account whether the birthdate given in the standard recode file is imputed, to determine whether imputation has a role in the displacement (there is no evidence that it does have a role). We also do not take into account the relationship between the household respondent—who provides the initial stated age for everyone in the household—and the person whose age is stated. The household respondent typically reappears in the survey of women or the survey of men. The household respondent is often the mother of children in the household who reappear in the birth histories. In-depth exploration of such effects is not possible in this report.

2.5 Data and Subpopulations

Four subpopulations are described: (1) children age 0-4; (2) children age 5-14; (3) women age 15-49; and (4) men age 15-49. For all subpopulations, the criterion for inclusion is that the person is in the PR file, which provides hv105, and also has a birthdate. Thus, for children age 0-4, no limitation is placed on whether the child's mother is in the household and no distinction is made between *de jure* and *de facto* residency status. Subpopulation 1 includes all children age 0-4 for whom hv105 and a date of birth are present in the household file.

[6] An alternative way to obtain MOB1 is as the p-value from a likelihood-ratio chi-square goodness-of-fit test of the following log-linear model for the frequencies n in the combinations of mob and moi: "poisson n i.mob i.moi". If there is displacement into interval 4, then much of the effect of "i.mob", which is apparent seasonality, is spurious and can be absorbed with the model "poisson n i.moi d," where d is a dummy variable that is 1 if mob<moi and 0 otherwise.

[7] Day 31 is dropped because it occurs only 7 times in a calendar year.

Subpopulation 2, children age 5-14, includes all children who are in both the household file and their mother's birth history (and are not in subpopulations 1, 3, or 4). The match requires b16 (line number in the household) in the birth history. For the most recent surveys, day of birth is included in the birth history.

Among women age 15-49 (subpopulation 3) and men age 15-49 (subpopulation 4), the match requires v003 or mv003 (line number in the household) in the women's file or the men's file, respectively. Year and month of birth are provided in these files, but not the day of birth. The age range for men is restricted to the age range for women for better comparability with women.

The procedure has been applied to all surveys conducted since 2008 (including 2008). Some results will be presented for all these surveys, although special attention will be given to two surveys to illustrate how the statistical model can guide the assessment of a single survey.

2.6 Implications of Birthdate Displacement for Anthropometry

If a child is reported as younger than he/she actually is, then the child is assigned HAZ and WAZ scores that are too high, and is less likely to be in the left tail of the distributions. That is, there is a risk that a child who is actually stunted or underweight will be considered to be of normal height or weight. This direction of misreporting will occur when the child's birthdate is later than it should be—for example, in time intervals 4, 5, and 6 when it should be in an earlier interval. Conversely, if the child is reported as older than he/she actually is, with a birthdate that is too early, then the child is *more* likely to be in the left tail. Age displacement should have no effect on the WHZ and wasting. Birthdate displacement will also tend to increase the dispersion of the HAZ and WAZ scores but not the weight-for-height (WHZ) score.

If the assigned birthdate is correct, then the levels of stunting and underweight should be approximately the same in all six time intervals described above. If there has been systematic displacement, we expect to observe a gradient such that stunting and underweight are more prevalent in categories 1 and 2 than in categories 3 and 4, and less prevalent in categories 5 and 6. Indeed, such a gradient would provide supporting evidence of systematic displacement.

2.7 Application of the model for birthdate displacement

The methods to identify potential displacement of birthdate by year, month, and day will be illustrated with two surveys. The first—the Albania 2008-09 survey—shows little evidence of displacement, and the second—the Sierra Leone 2013 survey—shows a great deal.

Example 1: Children age 0-4 in the Albania 2008-09 DHS survey

Fieldwork for the Albania 2008-09 survey extended from October 20, 2008 to April 26, 2009, a duration of almost exactly 6 months. This survey was unusual in that year, month, and day of birth were collected for all persons in the household survey. Its inclusion as an example serves primarily as confirmation of the various indicators This analysis is limited to subpopulation 1, the 1,605 children under age 5 in the household survey for whom the recode files include both stated age and birthdate. The classification of these children's birthdates into the six time intervals is as follows:

```
int1   int2   int3   int4   int5   int6

   0      5    924    672      0      0
```

All except five children have birthdates that are consistent with their stated ages. All the discrepant cases are in interval 2. The observed numbers of cases in interval 3 and 4 are 924 and 672, respectively; the expected numbers, taking into account day of interview, are 904 and 692, respectively (after rounding; all calculations involve more decimal places than are shown). The expected frequencies are calculated to have the same sum as the observed frequencies. The analysis is summarized with the six indicators defined above:

YOB1=100*5/1605=0.0031 or 0.31%. This is the percentage of cases in intervals 1, 2, 5, or 6. It is negligible in magnitude.

YOB2=-0.12. This log odds describes the allocation to interval 4 versus interval 3, adjusted for the day of interview. Because it is slightly less than 0, there is a slight underallocation to interval 4. The exponentiated value, 0.89, is an adjusted odds of interval 4 versus interval 3.

YOB3=0.1166. The p-value of the test statistic for YOB2 rounds to 0.12. The deviation of YOB2 from 0 is not statistically significant, by any standard.

MOB1=0.33. This is the p-value for a chi-square test of the independence of month of birth and month of interview. This value does not indicate statistical significance, by any standard.

DOB1=6.28%. This is the index of dissimilarity for a comparison of the observed distribution across days 1-30 with a uniform distribution, ignoring year, month, and time interval. A total of 6.28% of cases would have to be shifted from overreported days to underreported days to achieve a uniform distribution. This is a relatively small percentage, which does not indicate a high level of displacement by day, and does not exceed the nominal threshold of 10%.

DOB2=0.0982. The p-value of the test statistic for DOB1 rounds to 0.10. It is not significant, by any standard.

Thus for children age 0-4 in the Albania 2008-09 survey, we have a very high level of consistency in the placement into time intervals and no evidence of reporting preferences for specific months and days, measured either by level or by statistical significance, results that would have been expected for this survey.

Example 2: Children age 0-4 in the Sierra Leone 2013 DHS survey

The surveys in Sierra Leone have often been cited for data quality concerns. The fieldwork for the 2013 survey was conducted from June 2 through November 3, 2013, a duration of almost exactly 5 months. The sample was much larger than the Albania 2008-09 sample, with 9,758 children under age 5 in the household survey with values of hv105 as well as day, month, and year of birth. The frequency distribution of these children across the six time intervals is as follows.

```
int1   int2   int3   int4   int5   int6

  92    239   2193   6455    604    175
```

About 89% of children were placed into intervals 3-4, which are consistent with their stated age. About 3% were in intervals 1-2 and 8% in intervals 5-6. The expected numbers of cases in intervals 3 and 4, calculated in such a way that the observed and expected totals are the same and taking the day of interview in account, are 3,532 and 5,116, respectively. The ratio of observed to expected is 0.62 in category 3 and 1.26 in category 4. There is immediate evidence of a bias toward the calendar year of birth, which is given by subtracting age from the year of the survey. The six indicators are:

YOB1=11.38%. The percentage of children that appear in the inconsistent categories 1, 2, 5, and 6 implies a relatively large level of displacement.

YOB2=0.73. This is the log odds of interval 4 versus interval 3, adjusted for day of interview. A value greater than 0 indicates that there are more cases in interval 4 than would have been expected. The exponentiated value of YOB2 is 2.08: the observed odds of placement into interval 4, relative to interval 3, are 2.08 times what would have been expected from the day of interview.

YOB3, the p-value of YOB2, is less than .0001. This is far beyond a nominal threshold of .001.

MOB1, the p-value for a test of independence between month of birth and month of interview is also less than .0001.

DOB1=12.26%. This is the percentage of cases that would have to be shifted from an overreported day to an underreported day to obtain a uniform distribution across days of birth 1-30. The value exceeds the nominal threshold of 10%.

DOB2, the p-value of DOB1, is less than .0001. The departure from a uniform distribution is highly significant.

The interpretation of the results for children age 0-4 in the Sierra Leone 2013 survey is that (a) there was substantial displacement out of the consistent intervals; (b) within the two consistent intervals (3 and 4), there was very substantial preference for the calendar year of birth that was year of survey minus stated years of age (interval 4); (c) the distribution of calendar month was significantly related to the month of interview; and (d) the distribution of calendar day of birth, within a month, was not uniform. This is new evidence regarding the quality of age and date reporting that goes beyond the more established indicators of incompleteness, heaping, and displacement.

3 FERTILITY

3.1 Introduction

Most indicators of data quality are distinctly different from the outcomes of substantive interest that are analyzed with DHS data. For example, two generic examples of data quality indicators described in Chapter 2 are incompleteness of age and digit preference in the reports of single years of age. However, providing an accurate description of a country's age distribution is not the purpose of a DHS survey. The assumption, usually implicit rather than explicit, motivating the use of such indicators is that they are associated with the accuracy of the main outcomes of interest such as estimates of fertility, under-5 mortality, maternal mortality, contraceptive prevalence, and stunting.

The potential association between such data quality indicators and the accuracy of the main estimates could be based on the inference that a survey that is not very successful at obtaining accurate values of age will not be very successful at collecting the components of more complex indicators. Although vague, this kind of inference by analogy is often the unspoken justification for data quality analysis.

The actual link between the data quality indicators and the accuracy of the main outcomes is rarely articulated. It is possible that complex indicators such as fertility rates have low mathematical or statistical sensitivity to the incompleteness of age and digit preference. Future research may clarify the potential sensitivity. In the absence of such analysis, it could be argued that a more efficient strategy for understanding data quality would be to assess the main outcomes of interest first, and then, if there appear to be problems, trying to trace those problems to age heaping, for example.

This chapter will focus on one of the main outcomes—fertility. The same strategy could be extended to under-5, adult, or maternal mortality rates, which also have a reference period of time and are derived from retrospective histories.

Fertility rates are among the most important indicators generated by DHS surveys. They have many uses that range from assessing the impact of family planning programs to projecting future population growth. They are subject to many potential sources of error, although, as stated above, the degree of sensitivity to the different potential sources of error has not been worked out formally. DHS has produced several reports that describe strategies for analyzing fertility and for identifying and interpreting evidence of reporting errors. One report (MR11) focused on evidence of omission and displacement in the birth histories. Another report (MR12) focused on the rates themselves. This chapter modifies the strategy used in MR12 and in some specific survey reports, and illustrates its application. It is described as the survey overlap approach.

The essence of the survey overlap approach is to examine the correspondence between two successive surveys during a 5-year interval of overlapping birth histories. For a specific country, the two surveys are labeled in chronological sequence as Survey 1 and Survey 2. The more recent survey, the survey whose data quality is being assessed, is Survey 2. The rates from Survey 1 appeared in the report on that survey and the usual interest is in whether the rates from Survey 2, when they become available, indicate that fertility has been declining. The interpretation of the estimates in Survey 2 depends on whether the estimates from Survey 2 match those from Survey 1 during the interval of overlap. If the estimates do not match, the whole series from Survey 2 becomes suspicious.

Of course, if Survey 2 does not match Survey 1 during the interval of overlap, it is possible that the problem lies with Survey 1, or that both surveys are problematic. Even if they do agree, it is possible that one or both surveys are inaccurate. In general, if the surveys agree during the period of overlap, it seems likely that *changes* in fertility, at least, will be estimated correctly. Sometimes, the assessment of the quality of Survey 2 is revised again after the next survey (Survey 3) is completed.

3.2 Improving the Comparison of Successive Surveys

When successive surveys are available from the same country, the analyst can combine the surveys to construct a continuous trajectory of change in age-specific rates or the total fertility rate (TFR), as illustrated in MR12. Such pooling of surveys takes all estimates at face value and does not make adjustments. In this report, the focus is on the most recent survey, and on identifying evidence of a discontinuity with the previous survey.

Fertility rates for DHS surveys are normally calculated for years ago, also described as an interval of years before the survey. For example, if a woman was interviewed in August 2007, her contributions to the numerators and denominators of the 5-year fertility rates would be her births and exposure during the 60 months from August 2004 through July 2007, inclusive. (Births and exposure during the month of interview are excluded because the observation of that month is incomplete.) Because the fieldwork takes place over several months, usually 3-6 months, the reference period is slightly different for women interviewed in different months. The same kind of blurring characterizes estimates for earlier periods, such as 5-9 or 10-14 years before the survey

DHS defines the fertility reference periods as years ago in order to produce estimates from the latest survey that are as current as possible. This approach makes it more difficult to compare with estimates from other sources, such as other DHS surveys from the same country or other countries. Even if Survey 1 and Survey 2 are approximately 5 years apart, a reference period that is 0-4 years before Survey 1 and 5-9 years before Survey 2 will not coincide exactly.

In this chapter, the first modification of the survey overlap procedure includes redefining the reference period as a range of *calendar years* prior to the beginning of fieldwork for the first survey. This provides a more comparable overlap between the two surveys. For example, if the fieldwork for Survey 1 began during 2007, the interval for the comparison will be the 5 calendar years before Survey 1, or the calendar years 2002-06.

The second modification is to include in the comparison both *single-year and 5-year* intervals within the period of overlap. We will calculate the standard age-specific fertility rates and the TFR for each of the 5 calendar years during the selected interval and for the 5-year interval as a whole, using the birth histories in Survey 1 and Survey 2. Discrepancies between the estimates from the two surveys, measured with a combination of arithmetic differences and ratios, will be interpreted as differences in the quality of the data.

The third modification is the use of statistical methods. In contrast to other methods in this report, the estimates will take into account the survey designs (weights, clustering, and stratification). Consistency between the fertility estimates from Surveys 1 and 2 will be assessed by the two criteria used in other chapters: level and statistical significance. The level criterion, which is similar to the prevalence criterion for other data quality indicators, is based on whether a discrepancy between the two estimates exceeds a numerical threshold. Specifically, if the arithmetic difference between the TFRs from the two surveys, for the period of overlap, exceeds 0.2 births, the difference will be flagged.

To assess the statistical significance of the difference, the women's data from the two surveys, including the birth histories, are combined into a single file, with survey as a covariate that takes the values 1 or 2. The fertility rates can be calculated in different ways (see the Guide to DHS Statistics and MR12). The computational approach used here is based on poisson regression. For each woman with any exposure to age 15-49 in the period of overlap, a separate record is constructed for each 5-year age interval in which she has any exposure. Such a record specifies the years of exposure (E) to the combination of age and time, the number of births (b=0, 1, 2, etc.), and the appropriate value of survey (1 or 2). The statistical model is poisson regression, with b as the outcome, ln(E) as an offset, suppression of a constant term, adjustments for the survey design, and the inclusion of survey as a covariate. The coefficient of survey has a standard error, confidence interval, and p-value, all of which are produced by the poisson regression command.

A fourth modification of the usual approach is a restriction on the age range of the TFR. The fertility estimates from both surveys, but especially Survey 2, for the interval of overlap, will be affected by progressive censoring of the older ages. Because the upper age for eligibility is 49, an estimate of the age-specific fertility rate for age 45-49 for 0-4 years before a survey has reduced exposure, and an estimate for 5-9 years before the survey will have NO exposure to age 45-49 and reduced exposure to age 40-44. In this analysis, we allow the two surveys to be up to 7 years apart in order to generate more potential pairs of surveys. For these two reasons, the age range used for the TFR—and for the age-specific rates—will be 15-39 rather than 15-49. The estimates of the age-specific rates and TFR using the surveys in this analysis suggest that a TFR for age 15-39 is approximately 87% to 99% of a TFR for age 15-49. Results based on age 15-39 will be misleading only if Survey 1 and Survey 2, in the same country, have substantially different percentages of the usual TFR in the age range 15-39.

Comparisons described above are made with pairs of surveys in which the first year of data collection in Survey 2 was (a) the calendar year 2000 or later, and was (b) at most 7 years later than the first year of data collection in Survey 1, and was (c) the most recent survey satisfying conditions (a) and (b). As of the closing date for this analysis, 46 pairs of surveys were identified. The primary indicator of the level of correspondence between the two surveys in a pair is the difference in the estimates of the 5-year TFR for age 15-39, calculated as the Survey 2 estimate minus the Survey 1 estimate.

3.3 Overview of the Discrepancy in TFR Estimates

For 15 pairs of surveys, about one-third of the 46 pairs, the difference in estimates, when rounded to the nearest tenth of a child, is only 0.0 or 0.1. This is a surprisingly high level of correspondence. The countries at this level, with a difference ranging from -0.03 to +0.14, are Honduras, Indonesia, Jordan, Peru, Tanzania, Cambodia, Maldives, Timor-Leste, Colombia, Rwanda, Tajikistan, Namibia, Burundi, and the Dominican Republic (ordered by increasing magnitude of the difference, although for this level of agreement, differences in magnitude are not important). The difference rounds to 0.2 for another nine surveys; to 0.4, or less, for 34 pairs of surveys; and to 0.5 or more for 12 pairs of surveys.

The full distribution of differences is shown in Figure 3.1. The differences are almost always positive, so that the estimate from Survey 2 is higher than the estimate from Survey 1. The negative difference that is largest in magnitude is only -0.03, which is negligible. The 12 surveys with the largest arithmetic differences are those from Niger (with a difference of 1.86), Sierra Leone (1.48), Ethiopia (1.36), Madagascar (1.09), Benin (1.08), Burkina Faso (0.97), Guinea (0.78), Liberia (0.73), Pakistan (0.65), Cameroon (0.64), Nigeria (0.62), and Senegal (0.57).

The comparison can also be made in terms of ratios of TFRs. A difference of 0.2 has a different meaning if the TFR is in the vicinity of 3 or in the vicinity of 6, for example. Another reason for examining ratios is that the poisson-based estimates are calculated on a log scale. Figure 3.2 shows the distribution of the ratios of the Survey 2 estimates to the Survey 1 estimates. We next examine in Figure 3.3 the differences and the ratios together in a single scatter plot. Pairs of surveys are flagged if they have an arithmetic difference greater than 0.5, a ratio greater than 1.10, and are in the upper-right quadrant formed by the red lines in Figure 3.3. There are 12 such pairs, the ones already listed for just the criterion of the arithmetic difference. For all 12 pairs, the difference is highly significant.

Figure 3.1 Distribution of the differences between two estimates of the Total Fertility Rate for age 15-39 during the 5 calendar years before the first survey, using the most recent pair of surveys in 46 countries

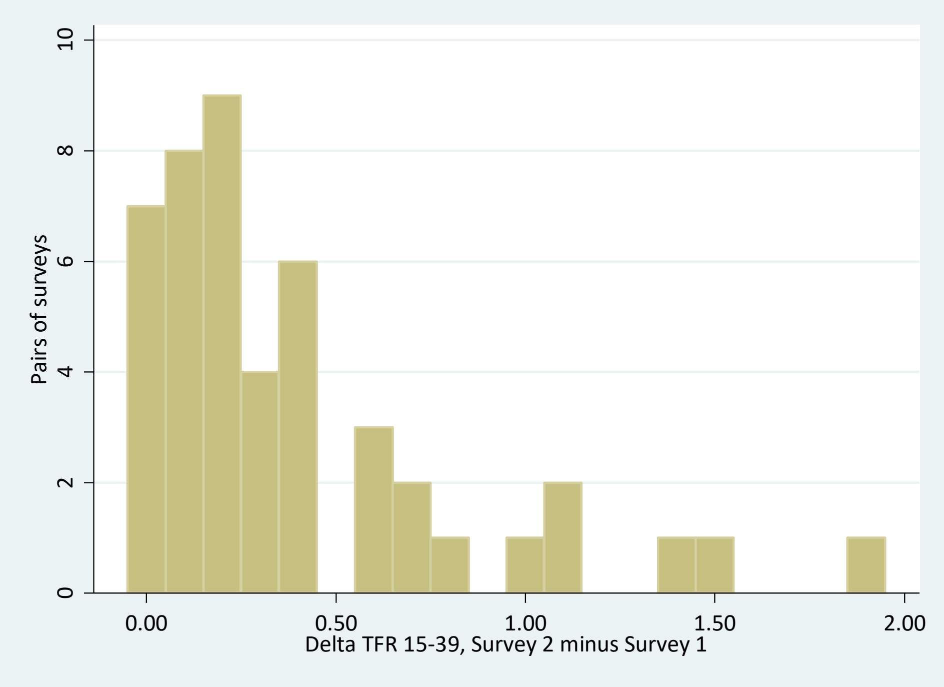


Figure 3.2 Distribution of the ratio of the two estimates of the Total Fertility Rate for age 15-39 during the 5 calendar years before the first survey, using the most recent pair of surveys in 46 countries

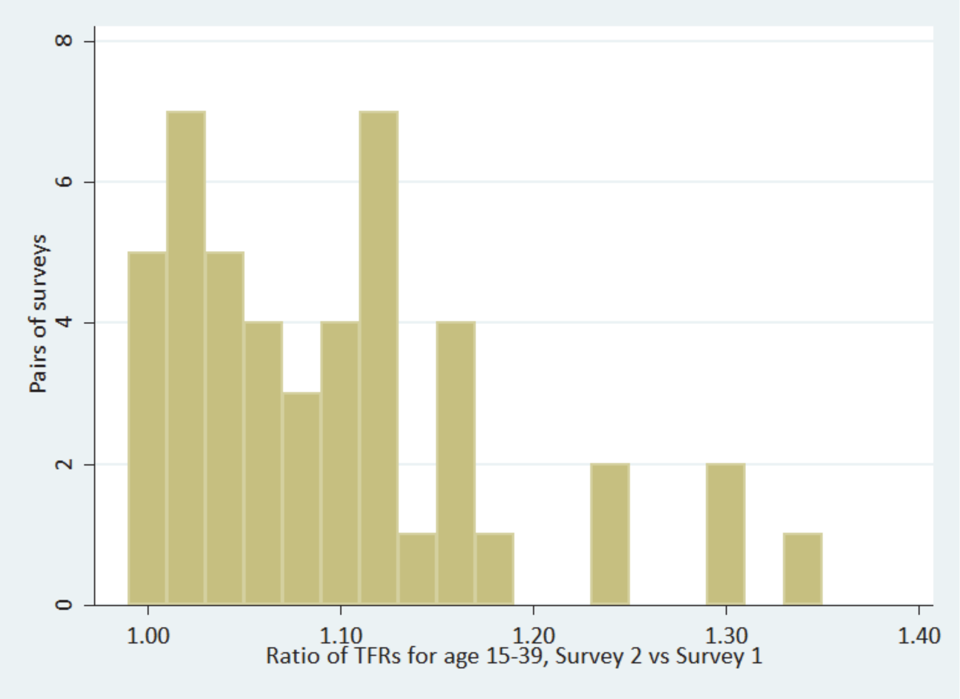
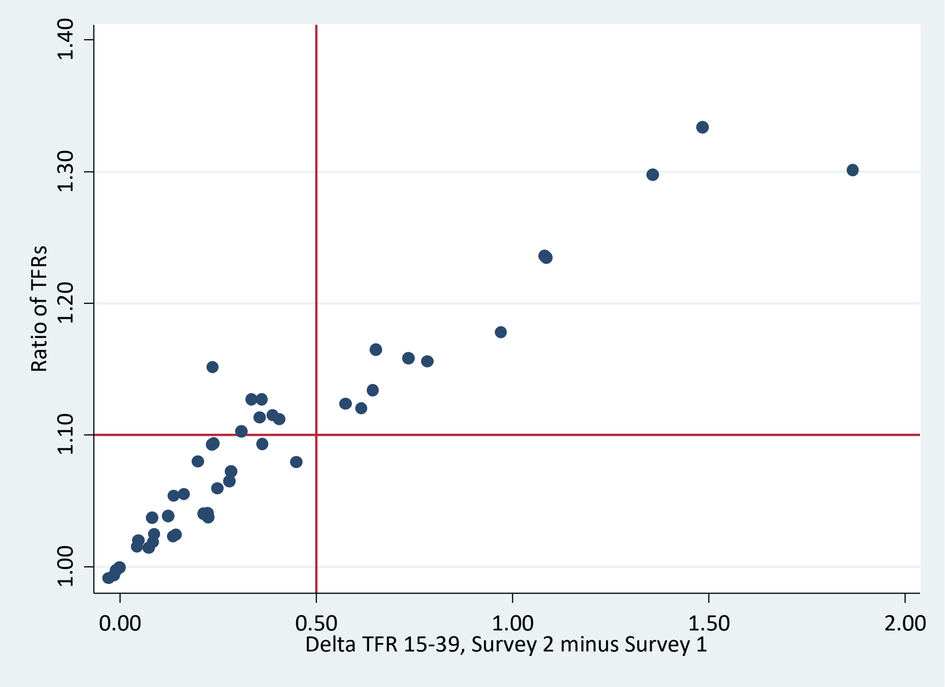


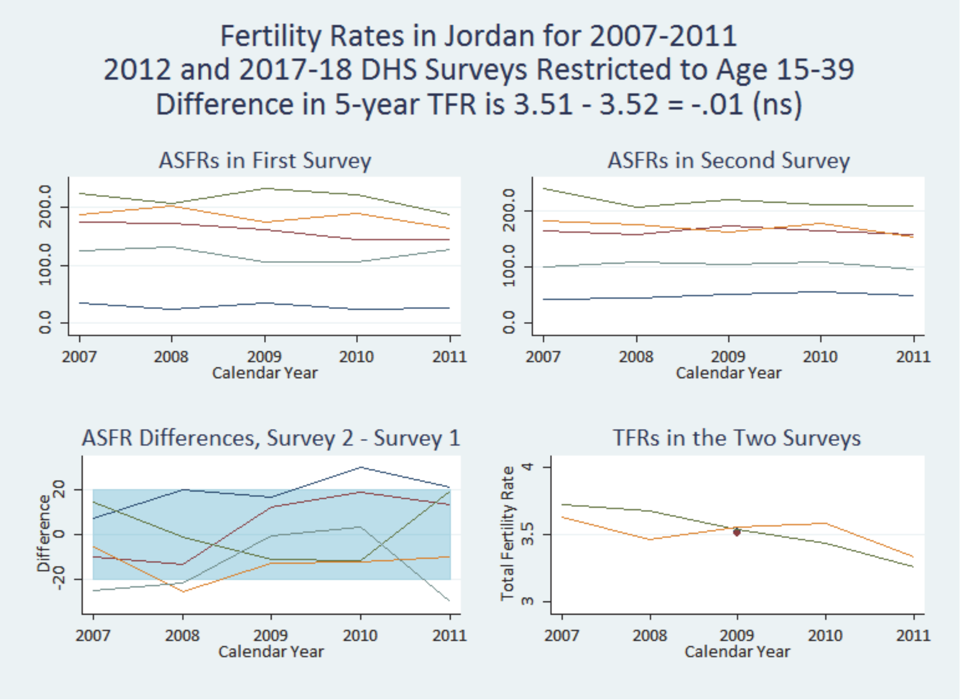
Figure 3.3 Scatterplot of the ratio and the difference of the two estimates of the Total Fertility Rate for age 15-39 during the 5 calendar years before the first survey, using the most recent pair of surveys in 46 countries



3.4 Results for Specific Surveys

Three of the 46 comparisons will be described in more detail. The first example, shown in Figure 3.4, is a comparison of the two most recent surveys in Jordan, which typifies pairs of surveys with very close agreement in their fertility estimates. This example will also serve to describe a set of four subfigures that are repeated for the other examples.

Figure 3.4 Age-specific and Total Fertility Rates for age 15-39 in the 2012 and 2017-18 surveys of Jordan, compared for the calendar years 2007-11



In Figure 3.4, all subfigures include single-year results. The single years are the 5 calendar years before 2012, the first year (and only year in this survey) of data collection in Survey 1. The ticks on the x-axis identify the midpoint of each calendar year, while the rate above it refers to the entire calendar year.

The top two subfigures show the five age-specific rates (15-19 through 35-39) in Survey 1 (left) and Survey 2 (right). The vertical axes of these two subfigures may vary from one comparison to another, although in this example the ticks correspond to levels of 0, 100, and 200 (births per 1,000 years of age-specific exposure). The legend is suppressed. The age group described with each line is not specified, although the colors for the age groups are the same in both subfigures.

The lower-left subfigure shows the differences between corresponding age-specific rates, calculated as the rate in Survey 2 minus the rate in Survey 1. A blue band ranging from -20 to +20 shows the tolerance region. The rate of 20 births per 1,000 years of exposure, when multiplied by 5, is equivalent to an impact of 0.1 children on the scale of the TFR. Differences in this range are of minor importance.

The lower-right subfigure has two lines that show the single-year trajectories of the TFR within the 5-year interval. The green line is based on Survey 1 and the orange line on Survey 2. In the middle of this subfigure there are two colored dots; in the Jordan example they are indistinguishable. The dots are placed above the middle year, in this example 2009, although they show the 5-year pooled TFRs from the two surveys. A blue dot refers to the estimate from Survey 1 and a red dot refers to Survey 2. In this example, those TFRs are 3.52 and 3.51, respectively. The difference, again expressed as the Survey 2 estimate minus the Survey 1 estimate, is 3.51-3.52=-0.01. The statistical significance of the difference in the 5-year TFRs is indicated in the figure, using ns for not significant and one, two, or three asterisks for p-values less than .05, .01, or .001, respectively.

The estimates in Figure 3.4 have an extremely high level of agreement. Virtually all differences in age-specific rates are within the tolerance level. The two lines in the lower-right subfigure are very close and the two dots are indistinguishable.

It is possible to read more in Figure 3.4 about the agreement between the two surveys during the period of overlap. Within the subfigures in the upper tier, the lowest line describes age 15-19. This age group has the lowest fertility in both surveys, although the estimates are higher in the second survey, especially for 2010. In the lower-left figure, the blue line represents the difference between the two estimates, and the increase in 2010 is a manifestation of the separation between the two blue lines in the upper subfigures.

All the Jordanian surveys are EMW surveys, that is, are limited to ever-married women, and all fertility estimates include an adjustment with all-women factors. The greatest sensitivity to the all-women factors is at age 15-19 because this is the age interval in which the fraction of women who are unmarried is greatest. The discrepancy in the fertility rates for age 15-19 could be traced to changes in the fertility of ever-married women in that age group or to changes in the proportion married, but the discrepancy is small.

A second comment about the lower-right subfigure in Figure 3.4 concerns the downward slope in the blue line for Survey 1, which contrasts with the virtually horizontal orange line for Survey 2. If we were to project Survey 1 into the years after 2011, we would expect to see a continuing decline in the TFR. The line for Survey 2 does not seem to be declining, but when Survey 2 is examined for the years 2012-16, it does show a declining TFR.

The second example includes two surveys in Niger, conducted in 2006 and 2012, with results shown in Figure 3.5. This pair of surveys had the greatest discrepancy across all 46 pairs in the analysis. The pooled 5-year estimate of the TFR for age 15-39 for calendar years 2001-05 was 6.21 in Survey 1 and 8.07 in Survey 2. The ratio of the second estimate to the first estimate is 8.07/6.21 = 1.30.

A comparison of the age-specific fertility rates in the upper two subfigures in Figure 3.5 shows clearly that all rates from Survey 2 were higher than the corresponding rates in Survey 1. Age-specific rates in Survey 2 approached 400 births per 1,000 years of exposure, which is implausibly high. The lower-left subfigure shows that the differences exceeded a tolerance level of 20, and were above the blue region of tolerance for all calendar years and every age-specific rate. In the lower-right subfigure, the separation between the dots representing the TFR 5-year estimates of 6.21 and 8.07 is very conspicuous. In that subfigure, the green line for Survey 1 and the orange line for Survey 2 show similar amounts of separation for the single-year estimates and the 5-year estimates. The next section of this chapter will explore the reasons why the estimates from Survey 2 are so much higher than those from Survey 1 for the interval of overlap.

Figure 3.5 Age-specific and Total Fertility Rates for age 15-39 in the 2006 and 2012 surveys of Niger, compared for the calendar years 2001-05

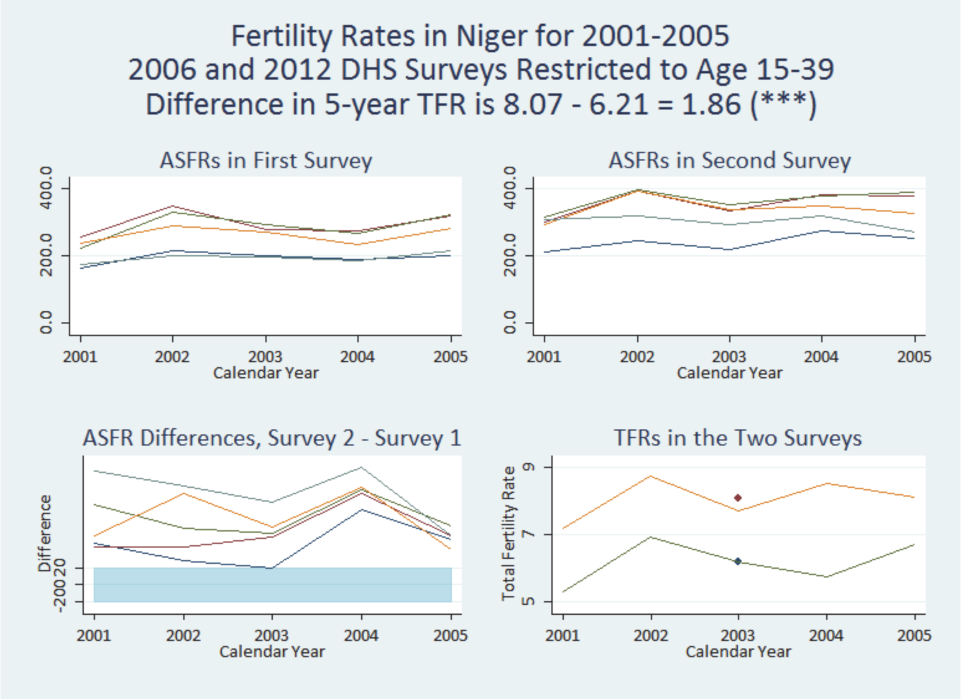
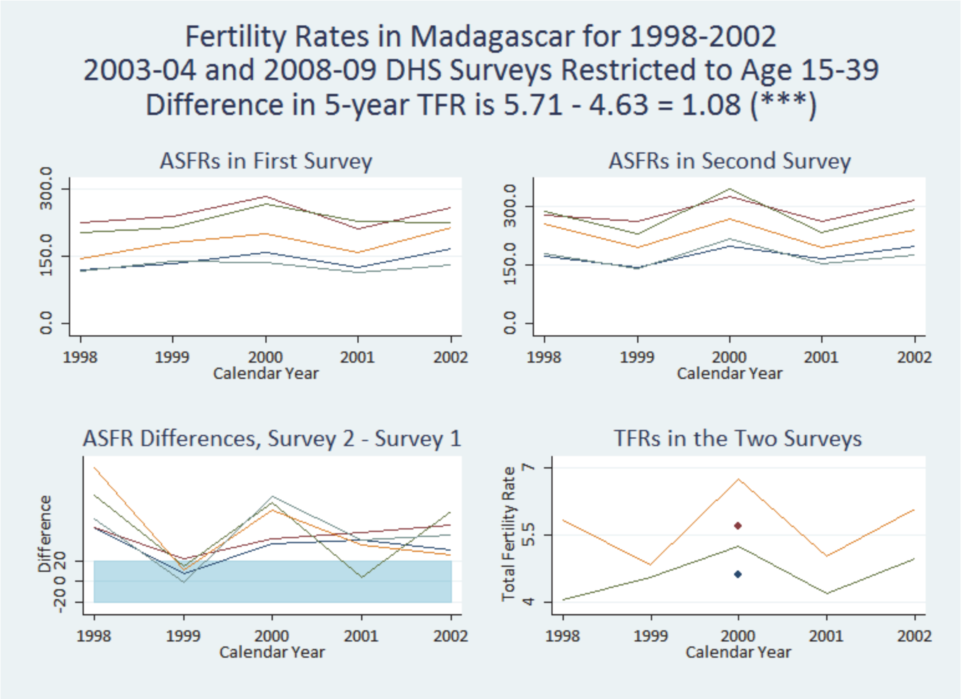


Figure 3.6 Age-specific and Total Fertility Rates for age 15-39 in the 2003-04 and 2008-09 surveys of Madagascar, compared for the calendar years 1998-2002



A third example of TFR comparisons is shown in Figure 3.6. The two surveys were conducted in Madagascar in 2003-04 and 2008-09. The TFR estimates were 4.63 and 5.71, respectively. The second estimate of the 5-year TFR for calendar years 1998-2002 is higher than the first estimate by 1.08 children. The pattern of deviations is the same as for Niger in Figure 3.5, although one additional source of error is very clear: a preference for calendar year 2000 as the reported year of birth. In Survey 1, all of the single-year age-specific births, and the single-year TFR, show an increase in 2000. In the second survey, which was 5 years later, the increase was even more pronounced, which demonstrates that with a longer time interval for recall, the greater the chance that a rounded response will be given. If the uptick in 2000 had the same magnitude in both surveys, then one could argue that perhaps some respondents were purposely timing their childbearing to be in 2000. The increased concentration on 2000 in the second survey clearly signifies that this is the result of number preference. Digit preference is well known for age, but here it is observed for year of birth. The excess for 2000 appears to be drawn virtually entirely from 1999 and 2001. Any pooling of years that included all three years (1999, 2000, and 2001), such as the 5-year TFRs for 1998-2002, is unaffected by the preference for 2000.

In comparing Figure 3.4 (for Jordan), Figure 3.5 (for Niger), and Figure 3.6 (for Madagascar), it is important to note that the vertical scales are not the same. For example, the lower-right subfigures show a range in TFR from 3 to 4, 5 to 9, and 4 to 7, respectively.

3.5 Potential Reasons for Differences between the Two Estimates

All 12 of the survey pairs that have a substantially higher TFR estimate for the reference period from Survey 2 than from Survey 1, although not presented here, have patterns similar to Niger in Figure 3.5, although they are less pronounced. We will briefly explore potential reasons for the differences.

Omission of births or misclassification of stillbirths

It is likely that in some contexts, some live births that resulted in very early deaths are omitted from the birth history (they are interpreted as stillbirths) or some stillbirths are misclassified as early neonatal deaths. This would only affect the comparison of surveys if such misclassification is related to elapsed time since the event and/or the likelihood is different in the two surveys. We observe that the fertility rates are higher in the reference period when estimated with Survey 2 than with Survey 1, which suggests a greater undercount of births in Survey 1 than Survey 2. This possibility is worrying but is not likely, because two successive surveys in the same country are generally similar in terms of factors such as training, supervision, and the implementing agency. Nevertheless, an undercount is a possibility.

Displacement of birthdates across the boundary for the health questions

Fieldworkers can avoid the health questions about children born in the previous 5 years (the interval since, and including, January of the fifth year before the beginning of fieldwork) by recording children as being somewhat older than they actually are. This shifts children out of (approximately) age 0-4 and into (approximately) age 5-9. If this happens in both Survey 1 and Survey 2, the effect is magnified in the type of comparison made in this chapter. For the reference period, the estimate from Survey 1 will be biased downwards, the estimate from Survey 2 will be biased upwards, and the difference between the two will be amplified. This may be the most likely explanation of the discrepancies.

Mortality of women

It is possible that the mortality of women can distort retrospective estimates. If women's mortality is positively related to their number of children, then a retrospective estimate would tend to be too low—the opposite of the pattern we see. It is unlikely that mortality over an interval of approximately 10 years, within the age range 15-49 (at the time of the survey), would be high enough to cause any discrepancy, particularly in the observed direction.

Heaping of women's ages on multiples of five

Women age 15-39 at the beginning of the 2001-05 reference period (for example, in Figure 3.5 for the two Niger surveys) would have been age 26-50 in 2012 (the time of the Survey 2 in Niger). DHS surveys only extend to age 49, so exposure to age 39—and births while age 39—in 2001 are reduced or omitted, although the effect should be negligible. Potential displacement across the age 15 boundary for eligibility for the women's interview and collection of the birth history is not relevant. Heaping on ages such as 30, 35, and 40 at the time of Survey 2 will translate into heaping on ages that are 11 years lower in 2001, 10 years lower in 2002, …, and 7 years lower in 2005. The effect on a rate should be negligible, because overreporting of women and exposure at those ages, in the denominators of rates, should be matched proportionately by overreporting of births at those ages, in the numerators of the rates.

Displacement of women's ages across age 15, the lower boundary for eligibility

Displacement across age 15 could have an effect on the estimated rates during the reference period from Survey 1, because the woman's age within the reference period is very close to her age at the time of the survey. The effect would be to slightly increase the age-specific rate for age 15-19 because the rate would be weighted toward the older part of the age group, which has higher fertility than the younger part of the age group. Displacement across age 15 in Survey 2 would have no effect on the estimates for the reference period.

We will not include more detail in this report on the deviations in specific surveys and the reasons behind the patterns. As stated in the introduction, a similar strategy can be followed to identify potential problems with other rates, such as the under-5 mortality rates.

4 VARIATIONS RELATED TO CHARACTERISTICS OF THE RESPONDENT

4.1 Introduction

Data quality depends critically on the execution of the survey, but it may also vary systematically according to characteristics of the respondents. Some of the problems associated with certain teams of fieldworkers may be attributable to the simple fact that the teams worked in geographic regions with respondents who cannot reliably provide their own or their children's birthdates or other basic information.

For each specific indicator of data quality, there is probably a combination of sources of variation. At one extreme, an outcome may be almost completely in the hands of the fieldworker. For example, anthropometric measurements are made by specially trained members of the team. Incorrect use of equipment is attributable to the fieldworker, and not the respondent. Other indicators of data quality may be affected by poor rapport between the interviewer and the respondent. For example, if the respondent and the interviewer have a large age difference, it is possible that the respondent will not be as forthcoming with answers to sensitive questions as she would be if the age difference were smaller. The quality of the response would then appear to be related to characteristics of both the respondent and the interviewer. At the other extreme, some data quality indicators may vary primarily by the respondent's characteristics, with minimal dependence on the fieldworkers.

To illustrate how a data quality outcome can be associated with respondent characteristics, this chapter considers the response rate for individual interviews and consent for HIV testing.

At the beginning of household data collection, respondents are asked to give their verbal consent to participate in the survey. Consent is also required to proceed with some topics or procedures such as HIV testing. Obtaining informed consent is consistent with respect for the rights of participants and is mandated by an Institutional Review Board approval process for every survey. Respondents can refuse to participate at the beginning or at any time during the interview. It is DHS policy to always respect the respondent's refusal.

Nonresponse for the individual interviews of adult women and men occurs when an eligible woman is identified during the household survey (with hv117=1) or an eligible man is identified (with hv118=1), but the individual does not appear in the recoded data files for the survey of women or the survey of men. In every main report, the response rate is given in Chapter 1 as a percentage. In this chapter, nonresponse for the individual interview occurs when the respondent refuses to be interviewed or when a respondent is not available, even after repeated visits. Another label for this outcome would be noncompletion of the interview. When considering HIV testing, the indicator is strictly limited to refusals.

In most DHS surveys, the nonresponse rate is low. Of DHS surveys since 2000, about 95% have a completion rate better than 90% for the women's survey and about 63% have a completion rate better than 90% for the men's survey. The sample weights are adjusted to account for nonresponse. Population-level estimates that use the weights (all DHS estimates are weighted) rest on an implicit assumption that the missing cases are not different from the interviewed/tested cases, in terms of background characteristics and outcomes of interest. However, if the level of nonresponse or missing cases is high, and the missing cases are systematically different from the interviewed/tested cases, there may be some bias in the estimates. For this reason, the levels of nonresponse and refusal can be viewed as indicators of data quality.

4.2 An Example of Respondent-level Variation

The strategy to assess the level of nonresponse/refusals and to identify respondent-level variation will be illustrated with the Malawi 2015-16 DHS survey. Two individual-level indicators are constructed.

First, for everyone in the household survey, nonresponse is coded 0 if an individual has hv117=1 or hv118=1; the code is changed to 1 if such a person does not appear in the file of women or the file of men (otherwise the code is NA). The source variable hv117 is coded 1 if the person is eligible for the individual interview of women, and hv118 is coded 1 if the person is eligible for the individual interview of men.

Second, for everyone in the household survey, the indicator of HIV refusal is coded 0 if an individual has code 1 or 2 for ha63 (women) or hb63 (men) and 1 if the person has code 3. HIV refusal is coded NA if the relevant source variable is NA. The source variable ha63 is the result of the HIV measurement for women and hb63 is the same for men. The codes of these two variables are 1: blood taken; 2: not present; 3: refused; 4: sample not tested/lost/damaged/insufficient; 5: not enough dried blood spot to complete protocol; 6: other. Codes 4, 5, and 6 are rare and are excluded from both the numerator and the denominator.

The 0/1 data quality indicator variables are then analyzed with logit regression, ignoring the weights or the sample design, as described in Chapter 1. The results can be used to produce bar graphs that readily identify covariates that are associated with nonresponse/refusals and to identify specific subpopulations with high levels.

Figures 4.1 and 4.2 include a total of eight subgraphs that describe the relationship between nonresponse and individual-level covariates. Figures 4.3 and 4.4 show the same for HIV refusals. Each figure includes a horizontal red line at the overall mean, which is 3.05% for nonresponse and 5.40% for HIV refusals. The eight covariates are:

Region of residence, a country-specific variable with only three categories for Malawi;

Type of place of residence (urban and rural);

Whether the person is the household respondent;

Household clustering, which describes any tendency for the data quality outcome to be 1 for another person in the household, given that it is 1 for at least one person;

Relation to household head, categorized as head, spouse, child, or other;

Sex of the person;

Age of the person, in 5-year age groups; and

Level of schooling, categorized as none, primary, secondary, and higher.

The figures include the pseudo R^2 value (the proportion of total deviance that is reduced by the covariate) and a statement of whether pseudo R^2 is significant with a p-value of .001 or less. This is a very stringent criterion for significance. The importance of a covariate is gauged in terms of both a high pseudo R^2 and a low p-value.

With nonresponse shown in Figures 4.1 and 4.2, most deviations from the mean level of 3.05% are small. The largest deviations are for the distinction between whether the individual was, or was not, the household respondent. Very few women or men who serve as the household respondent do not appear for an individual interview. Approximately 9% of the variation in the outcome is explained by this covariate. The next most important covariate is relation to head, which accounts for about 3% of the variation. The spouse of the head, usually a woman, has a much lower nonresponse rate than the head, usually a man. The categories of child and other have relatively high rates. The relevant children are eligible for the individual interview and therefore must be at least age 15 and still living with the parent or parents.

Two covariates have an R^2 value of .019, or about 2%. One of them is sex; the level of nonresponse for men is more than twice that for women. Household clustering is substantial but affects only households with two or more eligible respondents. If any one person in the household is a nonrespondent, the probability is greater for the other members. In other words, the probability of nonresponse is not independent for different household members.

Level of schooling or education is a highly significant covariate but only explains about 1% of the variation in nonresponse. The category with the highest level, by far, is for household members who have no formal schooling. The pattern is U-shaped; the next highest category is household members with postsecondary schooling.

Turning to HIV refusals, all covariates except age have very low p-values, but the largest pseudo-R^2 value is only about .15, or 1.5%, for clustering, and it only affects a small percentage of cases. Region and sex explain about 1% of the variation. HIV refusals are nearly twice as common for men as for women.

It is possible to extend this analysis to other covariates or to combinations—for example, including both sex and relation to head, perhaps in the form of a covariate that identifies all combinations of the two variables. Covariates that have high predictive value could be considered for inclusion as controls in analyses of interviewer effects. As mentioned earlier, interviewers should not be penalized for being assigned to areas or subpopulations with a high incidence of problematic responses. It could be useful to identify the subpopulations that have especially high levels of nonresponse or refusals, and the reasons for the nonresponse.

Figure 4.1 The percentage of eligible respondents who do not appear in the individual surveys of women and men, by selected covariates, in the Malawi 2015-16 DHS survey

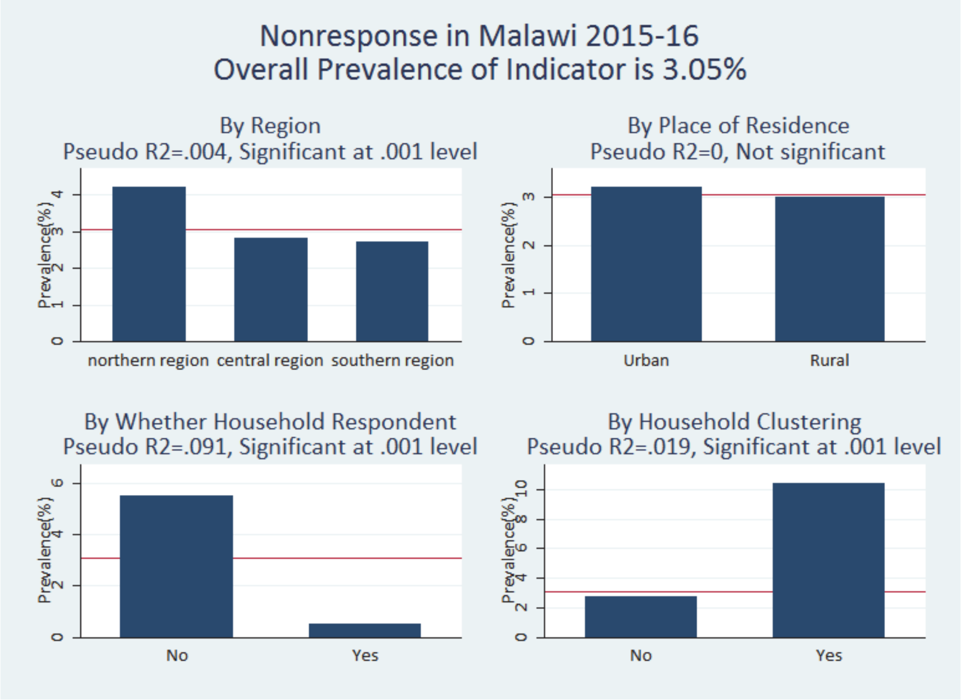


Figure 4.2 The percentage of eligible respondents who do not appear in the individual surveys of women and men, by selected covariates, in the Malawi 2015-16 DHS survey

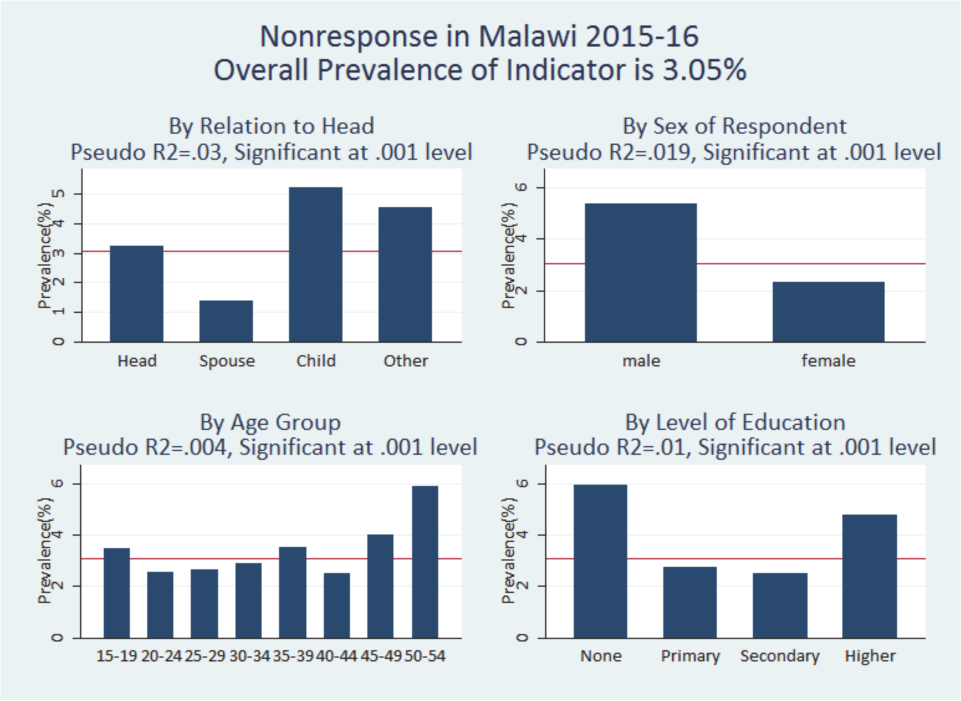


Figure 4.3 The percentage of eligible respondents who refused HIV testing, by selected covariates, in the Malawi 2015-16 DHS survey

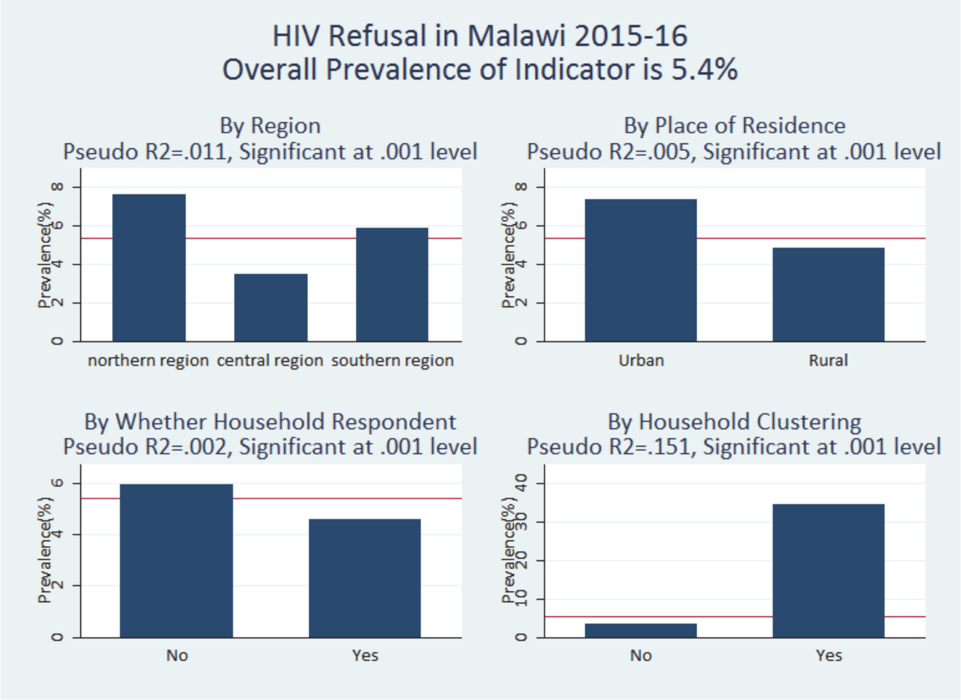
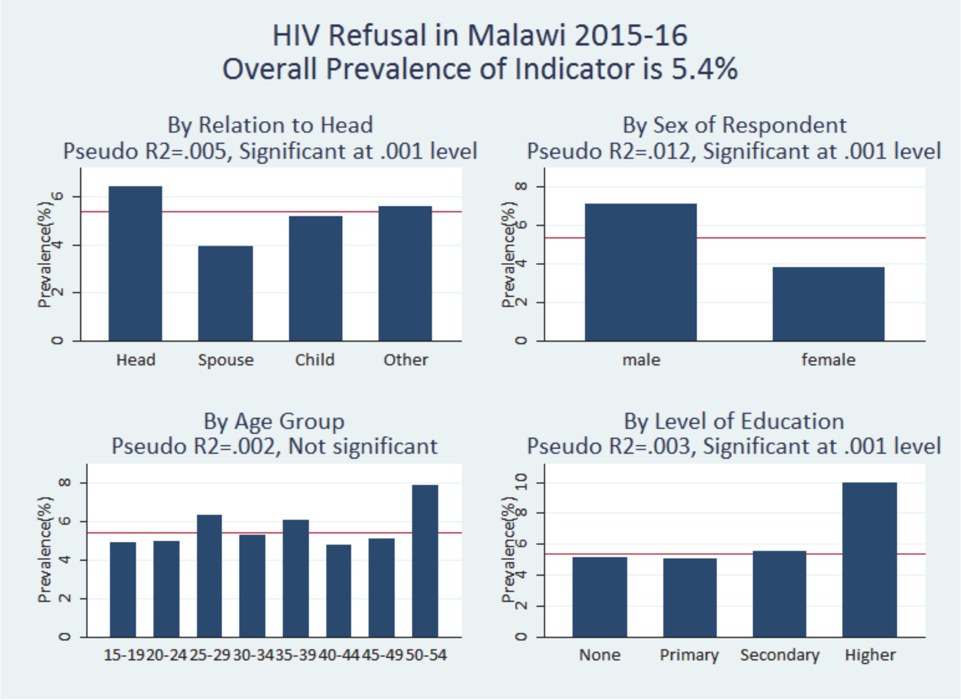


Figure 4.4 The percentage of eligible respondents who refused HIV testing, by selected covariates, in the Malawi 2015-16 DHS survey



5 VARIATION RELATED TO DURATION OF THE INTERVIEW

5.1 Introduction

The term process indicators refers to indicators based on characteristics of the interview that potentially reflect data quality. This chapter illustrates only selected uses of these process indicators and only scratches the surface of the many analytical possibilities. The increasing use of computer-assisted personal interviewing (CAPI) has produced reliable time stamps, for example, for the start and end of interviews. In the near future, DHS will have this information for blocks of questions and even for single questions. With CAPI, it should be more difficult for interviewers to deviate from normal practices without being detected.

Four process indicators are identified:

The length of the interview. The length or duration of the interview is measured in minutes, from the start time to the end time. The interval does not include the time that may be required to obtain anthropometric or biometric data. The analysis is limited to the household survey, although much of the interpretation carries over to the interview of individual women and men.

The length of the interview has two interpretations. First, length is a data quality outcome in its own right. The questionnaire, training, and management of a survey are designed to make data collection efficient. Logistical arrangements for the numbers of vehicles, drivers and all other staff, costing, and the sequencing and scheduling of sample clusters depend on accurate estimates of the average length of time required for the household and individual interviews. To remain on schedule, each team and each interviewer should conduct a specified number of interviews each day.

Very short interviews can be problematic. The interviewer should be efficient but should also maintain rapport with the household informant or respondent and not rush through the questionnaire Very long interviews are also undesirable because of the likelihood of respondent and interviewer fatigue. Thus, the prevalence of very short or very long interviews can be used directly as a data quality indicator.

If a short interview is undesirable, it should be possible to identify an empirical association with other indicators of poor data quality. This suggests that the duration of interview is an intervening variable or mechanism that affects other indicators. For example, we expect that very short interviews would result in cursory estimates of ages, omissions of births, inaccurate statements of children's ages, incomplete recall of contraceptive methods used during the past 5 years, and loss of quality with sensitive questions. Very long interviews may result in boredom or disengagement that increases the frequency of don't know, not applicable, or repetitive responses.[8]

The position of the interview, from the beginning to end of fieldwork, in the entire survey. During fieldwork, it is expected that all fieldworkers will become more efficient, as they become more familiar and

[8] We do not look at the time interval from the end of one interview to the beginning of the next, which partly measures the efficiency of moving from one household to the next or from one cluster to the next.

comfortable with the questionnaires, procedures, and teamwork. Some of this efficiency will result in a reduction in the duration of the interview. At the same time, there is a lower limit to the duration. The number of questions about household assets and the number per household member remain the same from the beginning to the end of fieldwork. Pressure to maintain a strict schedule that leads to the end of fieldwork may cause the later interviews to be rushed. The start times of the interviews within the entire survey are used to construct four quartiles, with the first quartile including the first 25% of interviews and the fourth quartile the final 25% of interviews.

The position of the interview, from the beginning to end of fieldwork, in each cluster. Within a specific cluster, there may be variation in the duration of the interview related to whether a household appears early or late during fieldwork. There may be a within-cluster gain in efficiency, or it is possible that an interview that takes place relatively late in the fieldwork for a specific cluster may be too short and suffer from a loss of data quality. The start times of the interviews within each cluster are used to construct four quartiles. The first quartile includes the first 25% of interviews in the cluster, while the fourth quartile includes the final 25% of interviews in the cluster.

The time of day when the interview began. It is possible that interviewers are under additional pressure to complete an interview quickly if the interview begins relatively late in the day. There is variation in the acceptability of conducting interviews early in the morning or in the evening, so time of day is not measured by the clock but by the relative position of the interview within the day. Using data in the entire survey on the start time of each interview, the interviews are divided into four quartiles. The first quartile includes the earliest 25% of interviews in the day, while the fourth quartile includes the 25% that are latest in the day.

The volume of information collected. The DHS questionnaire has evolved over successive phases to be progressively longer. There is some variation in the number of questions included in each survey and the number of questions across respondents. We use the number of household members and the number of non-missing responses to measure the volume of information.

This report will not attempt to establish the correspondence between the variables listed above and the prevalence of other data quality indicators. This chapter only considers the potential effect on the length of the household interview of the relative position of the interview in the entire course of fieldwork, relative position in the cluster's fieldwork, relative time of day, and the volume of information.

5.2 Data and Methods

The data for this chapter are limited to 22 surveys conducted in 2015 or later (and available during the preparation of this report) for which the start, end, and duration of the interview are coded as hv801, hv802, and hv803, respectively, in the household files. The specific surveys are listed below:

Afghanistan 2015; Albania 2017; Angola 2015-16; Armenia 2015-16; Benin 2017-18; Burundi 2016-17; Colombia 2015; Ethiopia 2016; Haiti 2016-17; Indonesia 2017; Jordan 2017-18; Malawi 2015-16; Maldives 2016-17; Nepal 2016; Pakistan 2017-18; Philippines 2017; South Africa 2016; Tajikistan 2017; Tanzania 2015-16; Timor-Leste 2016; Uganda 2016; and Zimbabwe 2015.

Both hv803 and v803, the length of interview in minutes for the household interview and women's interview, respectively, have the value 95 for durations of 95 minutes or longer. Very few household interviews have that code because it is very unlikely that a household interview would last more than 95 minutes. Analyses of the length of the women's interview should not use v803 because long interviews with women are more common. The length should be recalculated from v801 and v802 with another choice of upper limit because data entry errors can produce lengths that are implausible[9] (similarly for the mv variables in the men's data).

Analysis of the length of interview in this report is limited to the household survey, and more specifically to households for whom hv803 was less than 95 minutes and there was only one visit. It is difficult to interpret the start and end times when there was more than one visit. Some household interviews, in some surveys, have hv803 equal to 0 or NA. Neither code should ever appear. NA values are excluded. Values of 0 are retained, as coded, along with a few other implausibly short values.

In most of this report, the data quality indicators are individual-level binary outcomes. In this chapter, the outcome, the length of the household interview in minutes, is a household-level interval-level variable, hv803 in the HR file.[10] Regressions using this variable will be ordinary least squares (OLS) regressions, rather than logit regressions. Adjustments for design effects will not be made and all households will count equally. It would be possible to refine the model to prevent fitted values that are negative, but that will not be done.

DHS does not specify a lower boundary or minimum threshold length of interview. If it did—for example, if a household interview shorter than 5 minutes were considered problematic—then we could construct a binary variable that was 1 for a short interview and 0 for all others, and conduct a logit regression analysis very similar to other analyses. However, the length of the interview can vary considerably, and for legitimate reasons can be either unexpectedly long or unexpectedly short. In OD73, an assessment of the Niger 2017 DHS survey, one of the main reasons for recommending suppression of the survey was the high incidence of short interviews with women. However, apart from that analysis, there has been little systematic use of interview length as a factor in data quality analysis. For these reasons, this chapter will not use or even propose systematic criteria, and is simply descriptive and exploratory.

5.3 Effect of the Timing of Fieldwork on the Length of the Household Interview

Of the 22 surveys in this analysis, the Jordan 2017-18 survey had the shortest mean duration for the household survey at just 9.5 minutes. The survey also had the highest association between the length of the interview and the quartile of fieldwork. Seventeen percent of the variation was statistically explained by the quartile of fieldwork. Figure 5.1 displays results of this survey in a bar graph. The first group of four vertical bars shows the mean duration for the quartiles of position during the full span of fieldwork. The

[9] The times given by hv801 and hv802 use a 24-hour clock and have an "hours:minutes" format such that, for example, 14:25 is coded as 1425. It is necessary to break out the two digits for hours and the two digits for minutes when calculating the difference between hv801 and hv802; similarly for the coding of times in the IR and MR files.

[10] One record per household can also be obtained by using the PR file and reducing it to the household member with hvidx (line number) equal to 1. However, because of the requirements of section 5.4, we only use the HR file.

second group refers to position within the fieldwork in specific clusters. The third group shows position within the day of fieldwork, while a horizontal red line illustrates the overall mean (9.5 minutes).

The Jordan survey had a steady decline in interview length during fieldwork, a 28% drop from the 1st quartile to the 4th quartile, which was a decline from 11.7 minutes to 8.4 minutes. The mean duration during the last quartile of fieldwork was the shortest for any quartile in the 22 surveys, except for the 4th quartile in Albania, which was on average only 7.4 minutes. Figure 5.1 also shows that, by contrast, there was no variation across quartiles within clusters or within days. We interpret this pattern as evidence of improved efficiency during the course of fieldwork.

Figure 5.1 Mean duration of household interviews within quartiles of overall fieldwork, fieldwork within clusters, and fieldwork within days, for the Jordan 2017-18 DHS survey

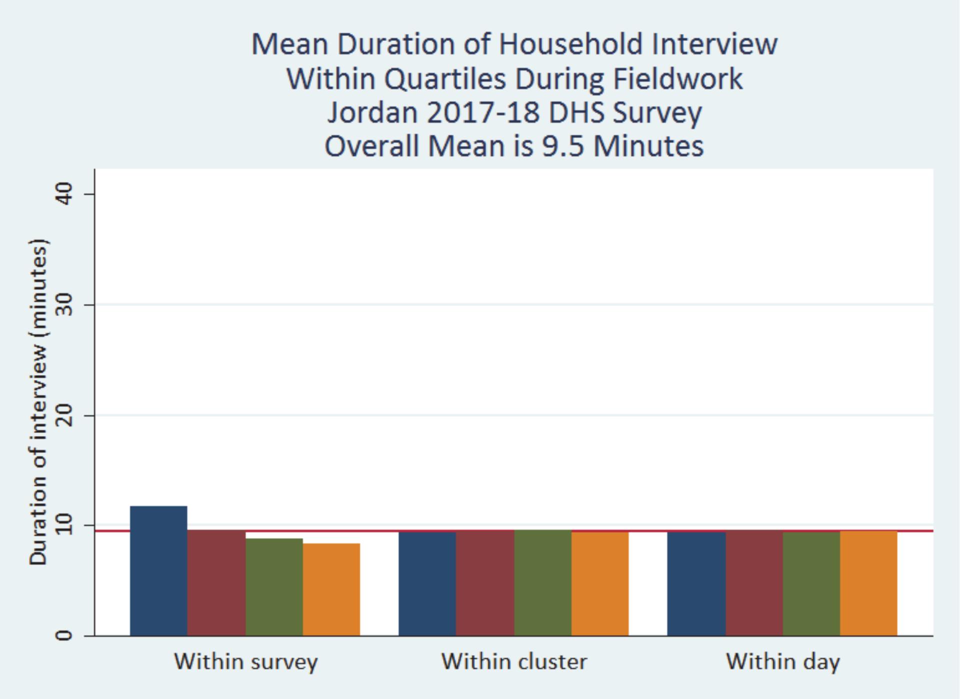


Figure 5.2 shows another pattern in the South Africa 2016 survey, in which all three measures of position show a progressive reduction in the duration of the interview. Within clusters and within days, the drop is primarily from the 3rd quartile to the 4th quartile, and mainly during the course of the day. This pattern suggests that there may have been pressure to complete the day's work, a circumstance that may be particularly likely if the team traveled to the cluster daily, rather than staying overnight in the cluster or nearby. A more thorough analysis would examine combinations of the three dimensions of time.

Figure 5.3 shows the pattern for the Angola 2016-16 survey, which had the sharpest decline in the duration of interview during the course of fieldwork, from 26.7 minutes in the 1st quartile to 14.7 minutes in the 4th quartile, a decline of 45%. This sharp drop may be due to a combination of improved efficiency and pressure to complete the survey. There was progressive decline in duration within clusters and within days, but this decline is steadier and less suggestive of time pressure than the pattern for South Africa in Figure 5.2.

Figure 5.2 Mean duration of household interviews within quartiles of overall fieldwork, fieldwork within clusters, and fieldwork within days, for the South Africa 2016 DHS survey

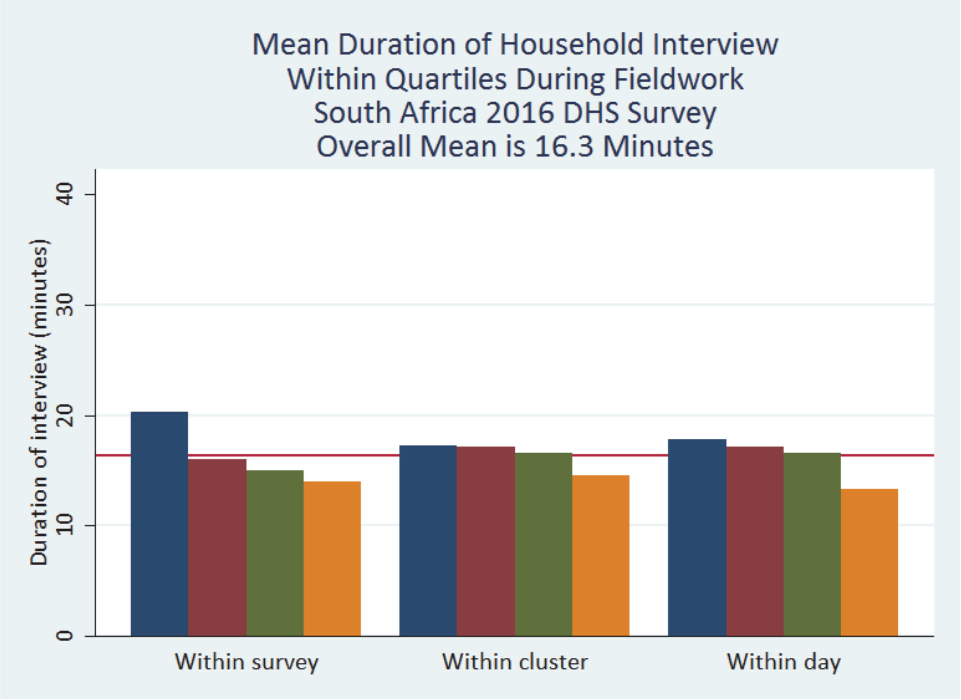
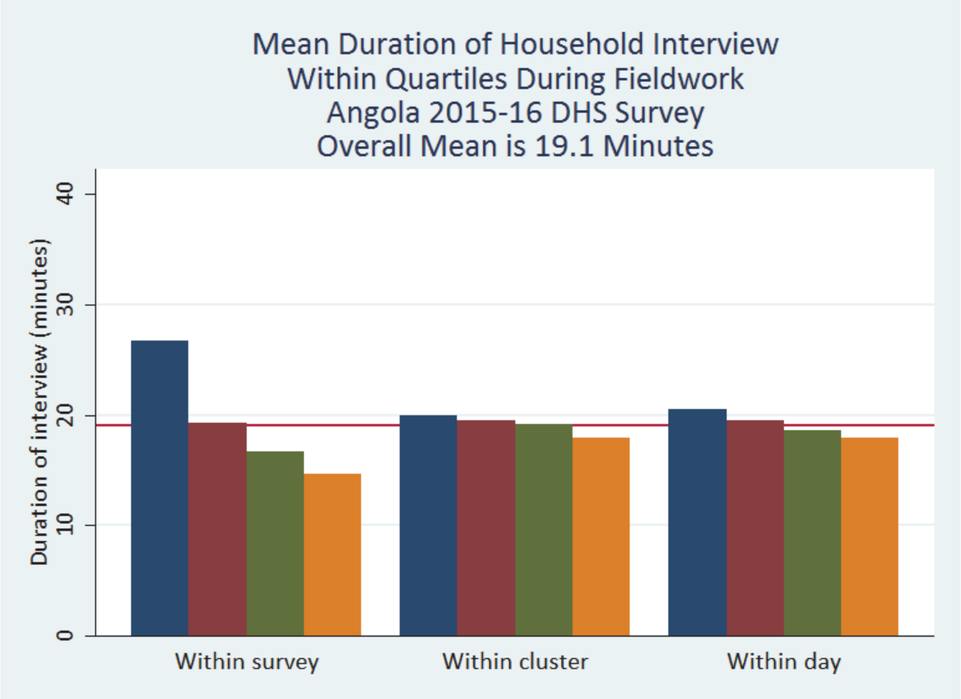


Figure 5.3 Mean duration of household interviews within quartiles of overall fieldwork, fieldwork within clusters, and fieldwork within days, for the Angola 2015-16 DHS survey



The Albania 2017 and the Maldives 2016-17 surveys are the only other surveys with a decline of more than 40% from the 1st to the 4th quartile of fieldwork. Most surveys had at least a 20% reduction, but two surveys showed an *increase* in duration over time—Afghanistan 2015 (7.7% increase) and Pakistan 2017-18 (2.4% increase). It is not suggested that either extreme in the evolution of the length of the interview is problematic. However, whenever an outcome is seriously atypical, such as a reduction of more than 40% in the length of the interview, or at the other extreme, an increase in the length of the interview, more exploration of the causes and implications is desirable.

5.4 The Relationship of the Length of the Household Interview to Household Size and Number of Items, across Surveys

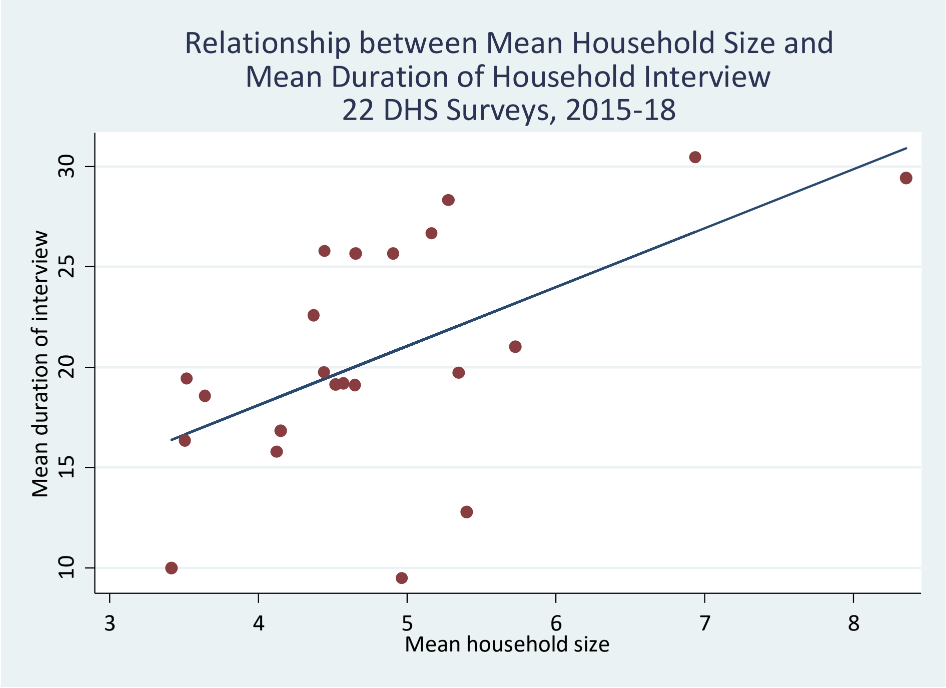
It would be expected that the length of the interview includes relatively fixed components of time, typically occurring mostly at the beginning and the end, plus a variable component of time that is roughly proportional to the number of household members and/or the number of items collected. Some household characteristics are coded automatically, such as region of residence and whether the area is urban or rural, and require virtually no time. These are the variables with names that begin with hv0, including hv009, the number of people in the household, regardless of whether they are *de facto* or *de jure*. Across the 22 surveys in this chapter, the two surveys with the longest mean duration of household interview are the two with the largest mean number of household residents. The Indonesia survey has a mean household size of 8.4 and a mean duration of 29.4 minutes, while the Pakistan survey has a mean household size of 6.9 and a mean duration of 30.5 minutes. The survey with the smallest mean household size, 3.4, is the Albania survey, which has the second shortest mean duration, 10.0 minutes. There is clearly a macro-level relationship between household size and the length of the household interview.

Figure 5.4 is a scatterplot that shows the 22 combinations of mean duration of household interview (y-axis) and the mean number of household members (x-axis) with a point for each survey. A line is fitted through the data. It has an intercept of 6.33 minutes, which is an estimate of the fixed or baseline duration, and a slope of 2.95 minutes, which implies an addition of almost 3 minutes for each person.[11] Note that both the vertical and horizontal axes of the figure are truncated.

Three surveys are substantially below the line, with a shorter observed mean duration than expected from the general pattern and the mean household size. In addition to the Albania survey, mentioned above, these are the surveys in Malawi (mean duration 9.5 minutes and mean household size 5.0 persons) and the Maldives (mean duration 12.8 minutes and mean household size 5.4 persons). Among the 19 other countries, the mean duration is at least 15.8 minutes, the baseline duration is 8.34 minutes, and the increment per person is 2.83 minutes.

[11] Every household has a minimum of one member, so an alternative statement of the minimum or base duration would be 6.33+1*2.95=9.28 minutes, but we prefer to identify the baseline with the intercept.

Figure 5.4 Scatterplot and fitted line showing the relationship between mean household size and mean duration of the household interview in 22 DHS surveys, 2015-18



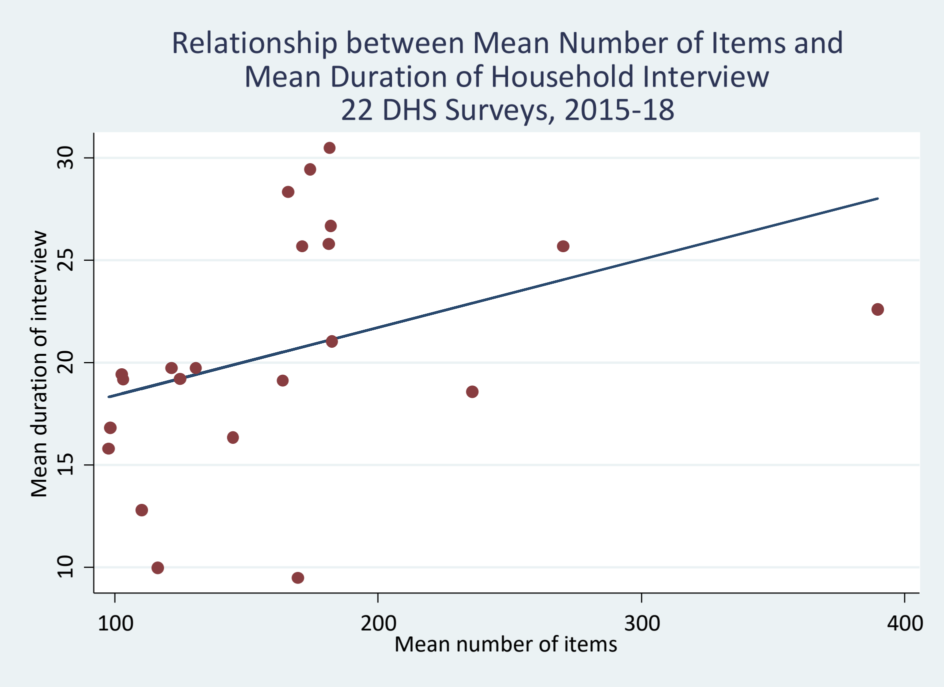
An alternative way to account for variation in the length of the interview is in terms of the number of items, or the amount of data collected. The majority of the relatively fixed component of time consists of questions about household possession/assets that range from whether the household has electricity, type of toilet facilities, and source of drinking water to the number of rooms and possession of items such as a motorbike or television. These are largely the questions used to construct the DHS wealth index, although many are also used in their own right. It is possible that interviewers can move quickly through some household-level questions by observing rather than explicitly asking. These are variables with names that begin with hv2, as well as some of the sh variables.

In most DHS surveys, there is a set of 12 to 20 items that must be collected for every person after all persons have been entered on a list. These include variables such as age, sex, the information to establish *de facto/de jure* residence status, marital status, education level, relationship to head, and eligibility for the interview of women or men. For children age 0-17 in most surveys, there are questions about the survival status of the mother and father and, if they are alive, whether they are in the same household as the child, and if they are in the household, who they are (as listed by line number). Ideally, when obtaining ages of household members, especially children, the interviewer will engage in some probing or use of a local calendar. Questions about school attendance or completed level of education are not asked about children who are younger than age 5. Survey-specific questions, for example, about health expenditures, are usually only asked about one person or just a few household members. Some questions may only be asked in a subsample of households. Thus, although the total number of items obtained about household members is not directly proportional to the number of household members, it is correlated with that number.

Variable names beginning with hv1 refer to standard data collected about individuals in the household. Variable names beginning with sh are survey-specific and may refer to the household, all individuals, or a subset of individuals. The household file may include some other blocks of variables, but those most related to the number of persons in the household are those with names beginning with hv1 and sh. For current purposes, we will not distinguish between sh variables that are household-level or that refer to all individuals or a subset. We construct a household-level count that is equal to the total number of hv1 and sh values that are not NA in the household. Across the 22 surveys included in this chapter, the mean count per household ranges from 98 in the Indonesia and Zimbabwe surveys to 390 in the Philippines survey.

Figure 5.5 is similar to Figure 5.4, but relates the mean duration of the interview to the mean number of items collected for individuals. The fitted line has an intercept (or base length of interview) of 15.07 minutes and a slope of .033, which implies that the interview length increases by about one minute for every 30 items. We believe that the intercept is too high and the slope too low, probably because many of the sh questions are asked about the households rather than about persons in the household. Moreover, the line is affected by a few outliers in the scatterplots, and without them the intercept would be lower and the slope would be larger. However, deviations from the fitted line are still informative. The Angola, Malawi, and Maldives surveys have a much shorter mean duration than expected, while several surveys also have substantially longer mean durations than would be expected. The three surveys with the longest duration (Pakistan, Afghanistan, and Benin) are farthest above the fitted line.

Figure 5.5 Scatterplot and fitted line showing the relationship between mean household size and mean number of items collected during the household interview in 22 DHS surveys, 2015-18



5.5 The Relationship of the Length of the Household Interview to Household Size and Number of Items, within Surveys

Most of the variation in interview length is within surveys rather than across surveys. In most surveys, some interviews are very short and some are very long. We now examine the degree to which the within-survey duration of interview is related to the number of persons and the number of items. If we can establish an expected duration, based on number of persons and/or number of items, we can then identify households with large deviations, particularly in the direction of being much shorter than expected.

First, in each survey, we regress duration of the household interview on the number of persons in the household. We extract four numbers from each regression: the intercept, which is interpreted as a fixed baseline length; the slope, which is the additional time required for each additional household member; the p-value for that slope; and R^2, which is interpreted as the proportion of variation in time that is explained (in a statistical sense) by the number of persons in the household. Second, the procedure is repeated for each survey using the number of items as the predictor of interview length.

When the duration of the household interview is regressed on the number of persons in the household, the slope (additional minutes per household member) is always positive and highly significant. The median is 2.08, just above two minutes, and in the range 1.27 to 2.88 minutes for all except three surveys with a smaller slope (Afghanistan, 0.51; Armenia, 0.57; and Indonesia, 0.79) and two surveys with a larger slope (Colombia, 3.31; and Benin, 3.29). In these same regressions, the intercept—interpreted as the base length of the interview—varies widely from 3.11 minutes (in the Jordan survey) to 25.16 minutes (in the Afghanistan survey).

Figure 5.6 Scatterplot showing the combinations of intercept and slope in the regressions of duration of household interview on household size in 22 DHS surveys, 2015-18

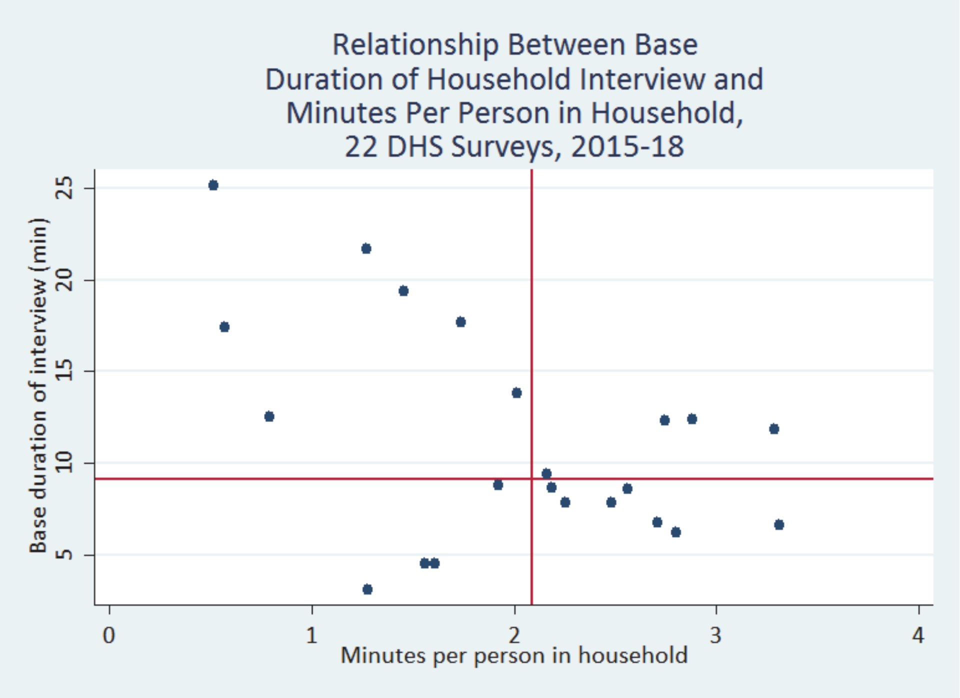


Figure 5.6 is a scatterplot that shows the base length of the household interview on the y axis (the intercept term in the regression) and the increment per person on the x axis (the slope in the regression). There is generally a negative relationship between these two, but we do not fit a line through the scatterplot. Rather, the figure includes two red lines at the median of each axis. Most surveys have a combination in the upper-left or lower-right quadrants. The Afghanistan survey, for example, is represented by the point that is highest on the vertical axis and lowest on the horizontal axis, because it featured household interviews with a long base length and a very low relationship to the number of people in the household.

We suggest that potentially problematic surveys are the ones in the lower-left quadrant. The interviews in these surveys have both a low base length and low sensitivity to household size. This is the quadrant where the survey would be located if the interviewers were moving too quickly and potentially skipping some content of the survey. Ignoring the Zimbabwe survey, which is in the lower-left quadrant but borderline, the surveys in this quadrant are the most recent surveys in Jordan, the Maldives, and Albania. These surveys are candidates for more detailed analysis.

Next, this procedure is applied to the number of non-missing items in the household interview, rather than the number of persons in the household, for each of the 22 surveys. The items are not questions but are the total number of non-missing responses to the hv1 and sh variables. This total includes some sh variables that differ from one survey to another, which do not actually contribute to the component of interview time that varies from household to another, but are fixed or relatively fixed in all interviews. A better implementation of the strategy used in this section would use a more accurate breakdown of the sh items, but for screening purposes, we classify all such items together.

Figure 5.7 Scatterplot showing the combinations of intercept and slope in the regressions of duration of household interview on number of items in 22 DHS surveys, 2015-18

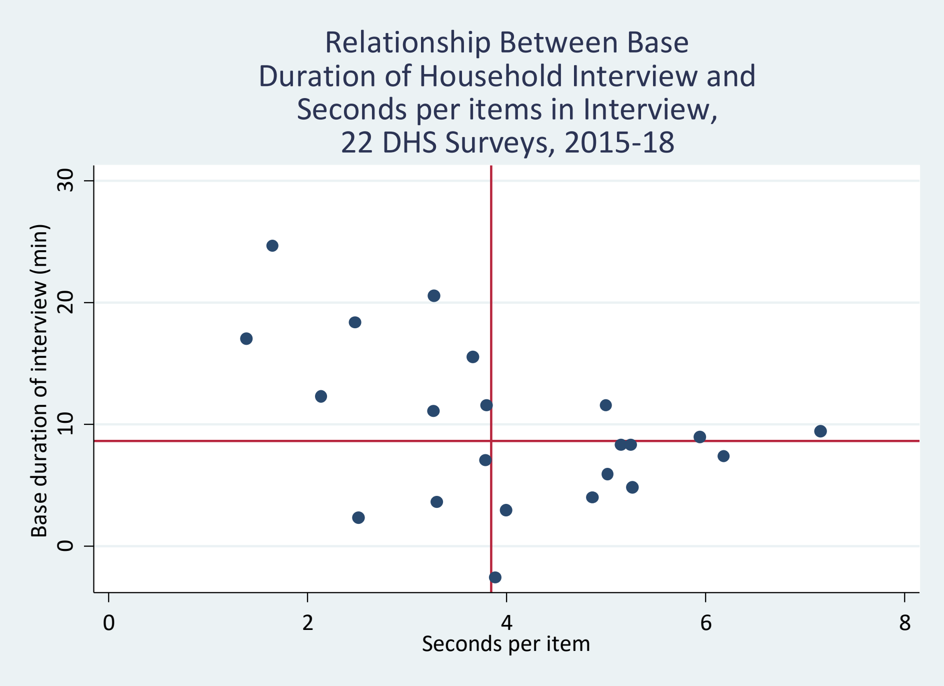


Figure 5.7 is analogous to Figure 5.6, except that it is based on regressions of interview length on the number of items, defined as above. The vertical axis is the intercept or base interview length from a regression and the horizontal axis is the slope, or additional interview time per item. The horizontal axis is measured in seconds rather than minutes. The median time per item is just under 4 seconds, and is almost entirely in a range from 2 to 6 seconds. This pace is much more plausible than what was calculated at the macro level. The base interview length calculated from the regression on items is not necessarily the same as the base calculated from the regression on persons, but the two are usually quite close. The median base durations in Figures 5.6 and 5.7 are almost identical, just under 10 minutes. In Figure 5.7, one survey has a negative fixed duration, which is likely based on a misallocation of the sh questions. A thorough survey-specific analysis would clarify the allocation.

As in Figure 5.6, most of the surveys in Figure 5.7 are in the upper-left or lower-right quadrants. We suggest that the potentially problematic surveys are those that are farthest into the lower-left quadrant—Jordan and Albania—and the survey with a negative fixed duration, the Philippines survey.

5.6 The Relationship of the Length of the Household Interview to Household Size and Number of Items in Specific Surveys

The surveys in Jordan and Albania stand out because of their presence in the lower-left quadrant of both Figure 5.6 and Figure 5.7. This outcome does not necessarily imply that these surveys are problematic, but it does suggest the importance of more analysis. We will focus on the Jordan survey. The goal is to determine the plausibility of the stated durations in terms of the number of persons and the number of items of data in the households. The strategy is described in terms of specific steps.

Step 1 Construct the process outcome. For each household, using the HR file, we construct the duration outcome as hv803, and drop interviews that required more than one visit (4,798 households, or 26.15% of the total 18,349 households). We also drop 57 households in which hv803 was coded as inconsistent[12] and 25 households in which hv803 was coded 95+. About 1% of households had a duration in the range of 26 to 93 minutes. To avoid an exaggerated effect from a few cases with relatively long durations, anything above 25 minutes was dropped.

Step 2 Clarify the structure of the other process variables. To establish an expected value for duration of interview, we use hv009 (number of household members), hv1count (the number of non-missing items per household interview that appear as variables with prefix hv1), and shcount (the number of non-missing items per household interview that appear as variables with prefix sh). These three variables are very highly correlated. If hv1count is regressed on hv009, the intercept can be ignored and a slope of 18.33 items per household member can be classified as part of hv1count. If shcount is regressed on hv009, we find an intercept of 21.18 items that are a constant for every household, and a slope of 11.68 items per household member that can be classified as part of shcount. The sum of 18.33 and 11.68 is 30.01. That is, after

[12] This was a CAPI survey and there should not have been any inconsistent cases. In non-CAPI surveys with inconsistent cases, many are due either to an obvious reversal of hv801 and hv802 (the start and end times) or to an obvious misuse of the 24-hour clock, such that noon was treated as 00:00 rather than 12:00.

rounding, 30 items on average are coded specifically for every member of the household, 18 are standard hv1 variables, and 12 are survey-specific sh variables.

The list of hv1 variables is as follows:

hv101	relationship to head
hv102	usual resident
hv103	slept last night
hv104	sex of household member
hv105	age of household member
hv106	highest educational level attained
hv107	highest year of education completed
hv108	education completed in single years
hv109	educational attainment
hv111	mother alive
hv112	mother's line number
hv113	father alive
hv114	father's line number
hv115	current marital status
hv116	currently, formerly, never married
hv117	eligibility for female interview
hv118	eligibility for male interview
hv120	children eligibility for height/weight and hemoglobin
hv121	member attended school during current school year
hv122	educational level during current school year
hv123	grade of education during current school year
hv124	education in single years - current school year
hv140	member has a birth certificate

As stated earlier, several questions apply only to certain age intervals. This survey omitted six standard questions related to schooling, hv110, and hv125-hv129. Most of the sh variables are in a block of approximately 20 variables that relate to outpatient health care and its cost in the previous six months, with a relatively small number of non-missing responses. Only 1,414 household members of a total of 93,347 household members have responses, but the questions were not restricted to a maximum of one respondent in a household, and the number of responses is approximately a linear function of household size.

An example of an sh variable that is specific to the context of this survey is sh07a, or nationality, which applies to every household member. This variable has six categories, plus a don’t know category, which we treat as a non-missing response.

Step 3 Develop a model to predict duration with the other process variables. A more elaborate model could include, for example, the quartile of fieldwork, quartile of fieldwork within the cluster, and quartile of the day and interview team, but for this illustration, the predicted or fitted values of the duration of the household interview will be based only on the number of household members, the number of hv1 variables, and the number of sh variables.

We do not provide full details on the model-building process, but provide an overview. Households are observed with a range of one to twenty members, although 99% have one to ten members. The model building focuses on that range and ignores the larger households. Within the range of one to ten, we check for the linearity of the effect of household size on the duration of the interview. The effect is almost perfectly linear. A regression of duration on hv009 has an intercept of 3.16 minutes and a slope of 1.20 minutes. The base is 3.16. However, since every household has at least one member, the mean length for a one-person household, both observed and fitted, is 3.16+1.20=4.36 minutes. The R^2 for this model is 0.33; a third of the variation in interview length is accounted for by number of persons. The fit is improved marginally by adding hv1count and shcount to the model, with R^2 increased to 0.37. All the predictors are extremely significant.

At this point, the principal findings for this survey are that (a) the mean duration for the smallest possible household, with just one person, is 4.36 minutes; (b) each additional person adds 1.20 minutes or 72 seconds; (c) each additional person adds an average of 30 items of information; and (d) the additional time per item of information is 72/30 = 2.4 seconds, on average. This is a highly accelerated rate of interviewing compared with other surveys, and could be investigated further.

Step 4 Investigate the cases with the most problematic deviations from the expected value. As a final step, the specific households with the shortest durations of the household interview, given the number of persons and items, would be identified. The analyst then needs to determine the reasons behind the inconsistency. It is possible, for example, that the start time or end time has obviously been coded incorrectly and that is why the duration is too short. If the duration is clearly implausibly short, there may be implications for the quality of the data recorded for the household. We will not pursue this last step within this report.

We emphasize that the reference to specific surveys with potentially problematic outcomes, such as the Jordan 2017-18 DHS survey in this chapter, is purely illustrative. This same survey appeared in Chapter 3 as an illustration of a favorable result for a different data quality indicator.

6 VARIATIONS DURING FIELDWORK

6.1 Introduction

Field checks are conducted routinely approximately every week during the course of every DHS survey. Survey managers are provided with tabulations, called field check tables, that describe a variety of indicators for every team for the interval since the beginning of fieldwork.

Most indicators of data quality that can be calculated from the completed data files can also be tracked during the course of fieldwork because the files include the date (year, month, and day) of interview. Following such a trajectory simulates the course of the fieldwork, from beginning to end.

The trajectories can be analyzed in at least the following four ways:

1. The data are collapsed by day, and sequenced chronologically by day. The aggregate of all days is the total survey.
2. The data are collapsed by day, and then cumulated, so that the results for day D describe all interviews up to and including day D. The cumulative total for the last day of fieldwork is the total survey.
3. The data are collapsed with clusters as units, and the clusters are sequenced chronologically by their mean date of interview. The aggregate of all clusters is the total survey.
4. The data are disaggregated by team, collapsed for each cluster visited by the team, and then sequenced by the mean date of interview of the cluster. This format produces a separate trajectory for each team.

This chapter will provide examples of each approach.

Any outcome—not just data quality outcomes—could be studied in terms of trajectories, but only a few outcomes other than those related to data quality would be likely to show variation over the course of the survey, perhaps reflecting seasonality of infectious diseases or reflecting the movement of fieldwork to successive geographic regions of the country.

There is evidence from field check tables that data quality tends to be poorest at the beginning of fieldwork, when interviewers and supervisors are establishing routines, and again toward the very end of fieldwork, when the interviewers anticipate returning home. Additional patterns could be hypothesized. It is possible that some type of misreporting—for example, backward displacement of birthdates—occurs early in a survey and is identified, corrected, and even *over*corrected after retraining. In this case, an analyst of the full data file, without looking at the pattern over time, could conclude that there was little or no displacement. It is also possible that a trajectory could reveal that some problem was not detected during fieldwork, and that there is a need for some modification of field checks.

6.2 Trajectories Collapsed and Sequenced by Day

To illustrate the day-to-day perspective on fieldwork, we look at the trajectory of the length or duration, in minutes, of the household survey. This approach pools the interviews that took place on the same calendar day, although they may have occurred in different clusters in very different parts of the country. The trajectories may include days during which few interviews occurred—perhaps none, if there was a gap in fieldwork. These days show more fluctuation simply because they are based on a smaller number of interviews. More detail on the length of the household interview was provided in Chapter 5.

We illustrate with four surveys:

Pakistan 2017-18, which had the longest mean length of interview, 30.5 minutes

Jordan 2017-18, which had the shortest mean length of interview, 9.5 minutes

Afghanistan 2015, which had the largest mean number of residents per household, 8.4

Colombia 2015, which had the smallest mean number of residents per household, 3.6.

The Afghanistan survey was also unusual because the length of the interview was not significantly related to variation in the number of residents.[13] In contrast, the Colombia survey had the highest observed sensitivity to the number of household residents. Each additional resident added 3.8 minutes to the length of the household interview. Apart from the Philippines survey, Colombia had the largest mean number of questions per additional resident—an average of 33. Thus, in Colombia, the household interviewer would typically take 3.8 minutes to record answers to 33 individual-level questions, which is an average rate of 9 responses per minute.[14] The rate increased over the duration of fieldwork.

Figure 6.1 shows the trajectory of the mean length of the household interview by day during fieldwork for the four surveys.[15] Each subfigure includes a smooth lowess line (with bandwidth 0.5). The Jordan and Colombia surveys are typical in that both show a steady reduction in the length of the interviews. The Afghanistan and Pakistan surveys are not typical because the interview length stayed at the same level throughout fieldwork. This pattern, although not the norm, does not necessarily imply a problem. The Afghanistan survey was the first complete DHS survey ever conducted in that country. The relatively long duration of the interview (in Pakistan as well as Afghanistan), and the steady maintenance of that long duration probably reflect a combination of the large mean household sizes and care taken during fieldwork.

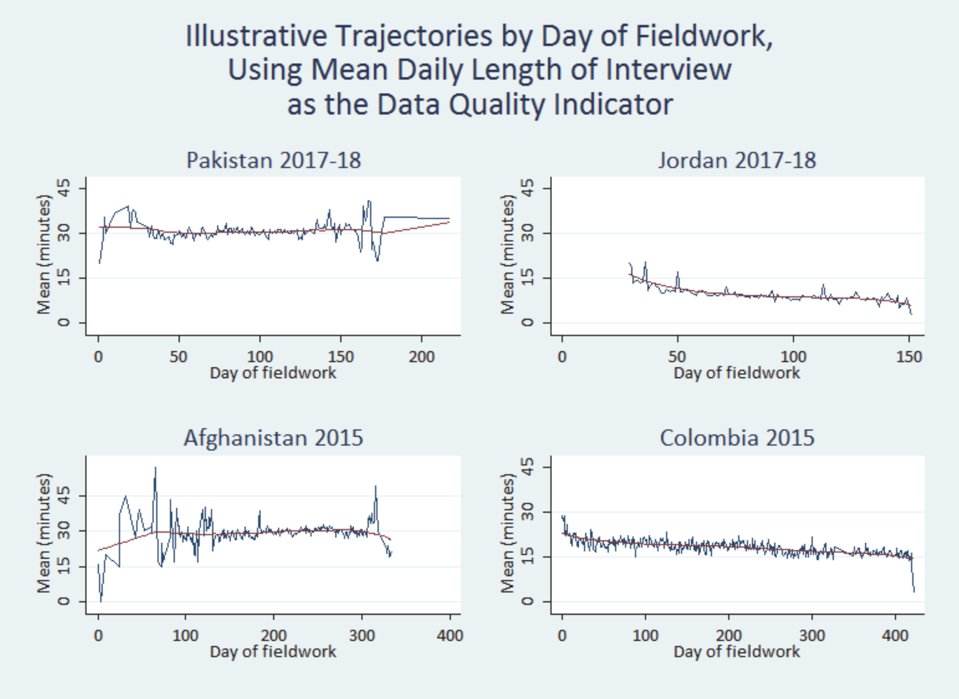
[13] The Pakistan 2017-18 file includes 14 households in which the day was early in the stated calendar year 2017, for which it is obvious that the calendar year was actually 2018. The year was recoded to 2018 for those households.

[14] In most analyses of DHS data quality, the recent surveys in Colombia can be used as a “gold standard.”

[15] In some surveys, the dates of interview in the recode files (hv008a) led to unexpected patterns. For example, in the Jordan survey, referring to the first day of interviews as day 1, there was one interview on day 1, one on day 7, three on day 29, five on day 30, and 114 on day 31. Taking the data at face value, there were only 10 interviews in the first 30 days of fieldwork. A similar pattern lies behind the other blank or erratic observations early or late in fieldwork for the other surveys in Figure 6.1.

As noted elsewhere, the short mean duration of the household interviews in the Jordan survey, combined with a substantial reduction (in percentage terms) from the beginning to the end of fieldwork, may suggest a deterioration of data quality over time and require further investigation. Most irregularities in a trajectory can be traced to days or weeks during which few interviews occurred, but some may reflect a problem encountered during fieldwork. The Afghanistan trajectory does show some irregularity, which is notable in the first third of the range of dates and near the end. The next step in a more thorough review would involve checking whether the incidence of displacement from age 15 to age 14, for example, appears more often when the duration of the interview is short or erratic.

Figure 6.1 Trajectory of the length of the household interview, by day during fieldwork, in the Pakistan 2017-18, Jordan 2017-18, Afghanistan 2015, and Colombia 2015 DHS surveys



6.3 Trajectories Collapsed by Day and Cumulated

Using the standard recode files, which include date of interview, every outcome in a DHS survey can potentially be calculated for each day of fieldwork, thus simulating what actually happened during fieldwork. An alternative way to follow fieldwork from day to day is to calculate an indicator using all observations up to, and including, each successive day of fieldwork. Such a procedure would show how the indicator converges to its final value, which is reached on the last day of fieldwork.[16]

A model for potential displacement of birthdates, described in Section 2.2, is well suited for this daily cumulative approach because the potential displacement into category 4 is related to the day of interview. The Angola 2008-09 and Sierra Leone 2013 surveys will again be used as examples. The only indicators

[16] For data quality indicators, the focus of this report, the calculations are unweighted. For indicators of substantive interest, the calculations would be weighted.

whose trajectories will be shown are YOB1 (the percentage of birthdates that are inconsistent with stated age) and YOB2 (the log odds of category 4 versus category 3). Trajectories will be provided for all four subpopulations defined in Section 2.2.

Seven computational steps are listed because they mimic what is alleged to occur during fieldwork. All the calculations are done with Stata.

Step 1. Every day of the survey is listed, from first day to last.

Step 2. The value of P is calculated for every day.

Step 3. The number of interviews on each day (cases appearing in the household rosters) is entered.

Step 4. Six separate columns give the numbers of cases falling into each of the six time intervals, on each day, based on a comparison of stated age with year, month, and day of birth. If day of birth is not provided, as with men and women age 15-49, the day is assumed to be 15. (In symbols, refer to these numbers as n1 through n6; their sum is n.)

Step 5. The expected allocations to categories 3 and 4 are calculated in such a way that the sum of the expected values matches the sum of the observed values. The expected number in category 3 is n3hat=(1-P)*(n3+n4) and the expected number in category 4 is n4hat=P*(n3+n4). Thus, for each day of fieldwork, n4hat/n3hat = P/(1-P).

Step 6. The numbers in the six columns of observed frequencies (for intervals 1 through 6, and for the sum of all intervals) are cumulated, giving the subtotals up to and including each day of fieldwork. The corresponding two columns of expected frequencies in intervals 3 and 4 (n3hat and n4hat) are also cumulated. The observed cumulative frequencies are cumn1 through cumn6 and cumn, and the expected frequencies are cumnhat3 and cumnhat4.

Step 7. The cumulative values of YOB1 and YOB2 are calculated for every day of fieldwork:

YOB1=100*(cumn1+cumn2+cumn5+cumn6)/cumn

YOB2= log[(cumn4/cumn3)/(cumn4hat/cumn3hat)]

After exponentiation, YOB2 gives the ratio of (a) the *observed* odds of category 4 versus category 3 to (b) the *expected* odds of 4 versus 3, up to and including each day of fieldwork. It can also be described as the adjusted odds of 4 versus 3. If there is no seasonality and the birthdates are assigned correctly, this ratio should be 1. If there is a preference for interval 4, the ratio will be greater than 1, with no theoretical upper limit. If there is a preference for interval 3, the ratio will be less than 1, with a theoretical lower bound of 0. The logarithmic transformation produces a symmetric measure that will be 0 if the allocation is correct, greater than 0 if there is a preference for interval 4, less than 0 if there is a preference for interval 3, and there is neither an upper bound nor a lower bound.

If, for example, on day 100 of fieldwork, YOB2 is greater than 0, then there is evidence that interval 4 has been overreported during the first 100 days. The overall results for each survey are given by the values of these measures on the last day of fieldwork and will match the values given in section 2.2.

Graphs will be produced for two surveys to show the trajectories of the four indicators during fieldwork. One graph shows YOB1 and a second shows YOB2. During the first days of data collection, the numbers are unstable. For this reason, the graphs do not include the first 10% of fieldwork. That starting point for the graphs is arbitrary. In large surveys, the indicators might stabilize earlier.

To make the graphs easier to interpret, each includes a blue rectangle that can be interpreted as a within-tolerance zone. Variations and fluctuations within this blue rectangle are judged not to be substantively significant, even if, after testing, they may be highly significant statistically. The vertical range of the blue rectangle for YOB1 extends from 0% to 5%. The vertical range of the blue rectangle for YOB2 extends from –log(1.25) to +log(1.25). Here 1.25 is set as a tolerance for an adjusted odds ratio. There is evidence that the maximum effect of true seasonality in birthdates is within this range.[17]

The vertical scale may differ from one figure to another. In general, if the vertical dimension or height of the blue rectangle is relatively large, then the range of the ratios is relatively small. If the height is small, the range of the ratios is large. The horizontal range of the blue rectangles is the range of the dates of fieldwork. A vertical line identifies January 1 if that date occurred within the duration of fieldwork.

Figures 6.2 and 6.3 show the trajectories for indicators YOB1 and YOB2, respectively, in the Albania 2008-09 DHS survey. As described earlier in this report, this survey showed very little evidence of birthdate displacement. Each figure includes four subfigures, for four subpopulations. In the subfigures and in the subsequent ones for the Sierra Leone 2013 survey, the vertical axes are not necessarily the same.

Figure 6.2 for the Albania 2008-09 survey shows that the level of YOB1 was below 5% and therefore within the blue tolerance box for the entire duration of the survey for all four subpopulations. The lowest level was for children age 0-4 and the highest for men age 15-49, although the differences among children age 5-14, women age 15-49, and men age 15-49 were negligible.

Figure 6.3 shows that the level of YOB2 was in the blue tolerance box for most of the survey for all subpopulations, and that the final value was very close to 0 for all subpopulations. However, for women and men, during the last month of 2008 (the first part of fieldwork), there was a conspicuous tendency to be placed in interval 4 rather than 3. In interviews after January 1, 2009, this displacement was largely avoided.

Figures 6.4 and 6.5 are analogous to Figures 6.2 and 6.3, respectively, for the Sierra Leone 2013 survey. As described earlier, this survey had strong evidence of birthdate displacement for children age 0-4. The figures show very similar patterns for the other three subpopulations, differing only by level.

During the course of this survey, the level of placement into intervals 1, 2, 5, and 6, as shown in Figure 6.4, steadily dropped. For all subpopulations, the final level was about half of the initial level. This type of

[17] Among the surveys included in a larger analysis, only four have a deficit of cases in time interval 4 relative to time interval 3, measured as a negative value of the logged odds ratio for category 4 versus category 3. The most extreme value is approximately –log(1.25). We assume that negative values only arise from seasonality. A range from –log(1.25) to +log(1.25) may thus exclude all influences of seasonality, but the range remains somewhat arbitrary.

displacement was initially quite high, and affected about 20% of children age 0-4 and about 30% of individuals in the other three subpopulations. By the final month or so, it was much less common.

Figure 6.5 shows that displacement into interval 4, compared with interval 3, was quite steady throughout the duration of fieldwork, although it was lower for children age 0-4 than for the other subpopulations. As seen previously, the final level of the adjusted log odds for children was 0.73 and the odds (the exponentiated log odds) were 2.08. For the other three subpopulations, the adjusted log odds ranged from 1.33 to 1.44, with the odds ranging from 3.77 to 4.24. For children age 0-4, placement into interval 4, rather than 3, was about two times greater than expected, and for all other respondents for whom a birthdate was calculated, about four times greater than expected. The steadiness in this pattern throughout fieldwork is clear. All fieldwork for this survey was done within a single calendar year, so there is no discontinuity at January 1. Some other surveys show an abrupt contrast between classification late in the first calendar year and early in the second calendar year.

Figure 6.2 Cumulative percentage of inconsistent birthdates (YOB1) in the Albania 2008-09 DHS survey

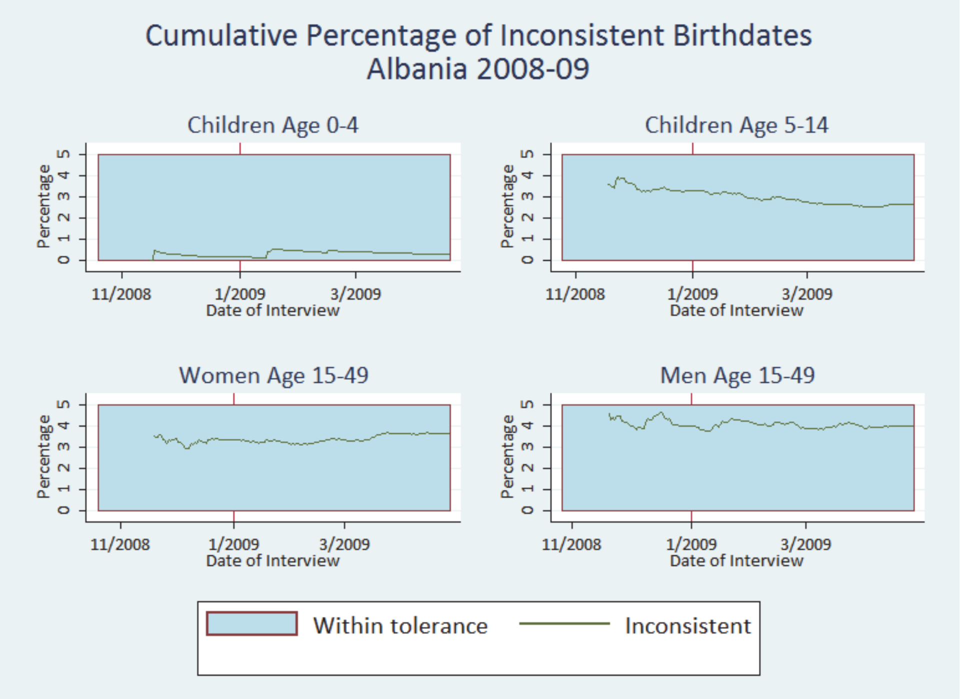


Figure 6.3 Log of the cumulative adjusted odds ratio for interval 4 versus interval 3 (YOB2) in the Albania 2008-09 DHS survey

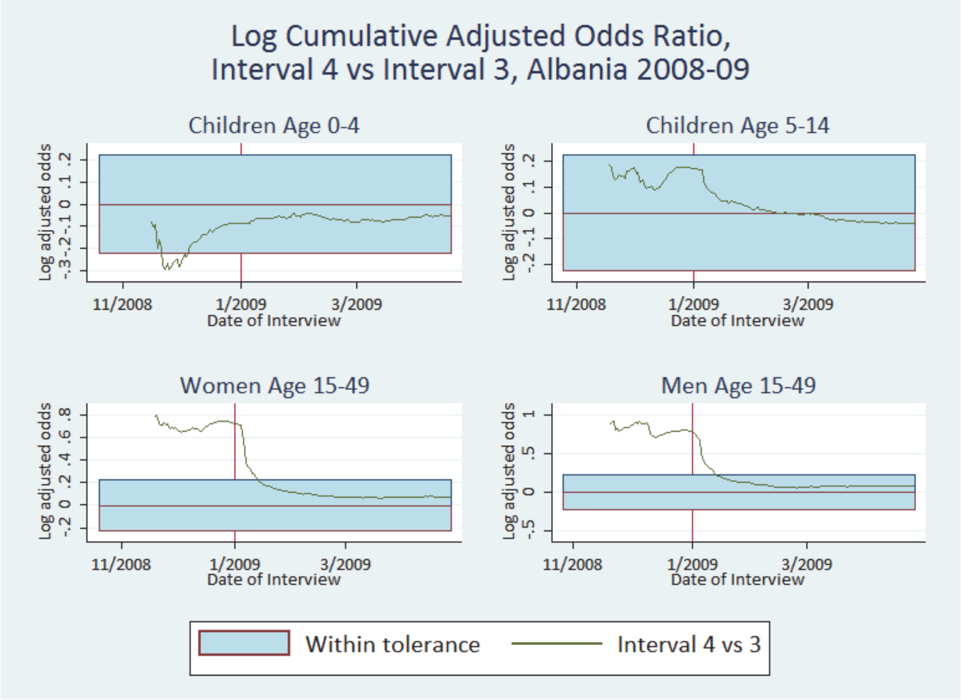


Figure 6.4 Cumulative percentage of inconsistent birthdates (YOB1) in the Sierra Leone 2013 DHS survey

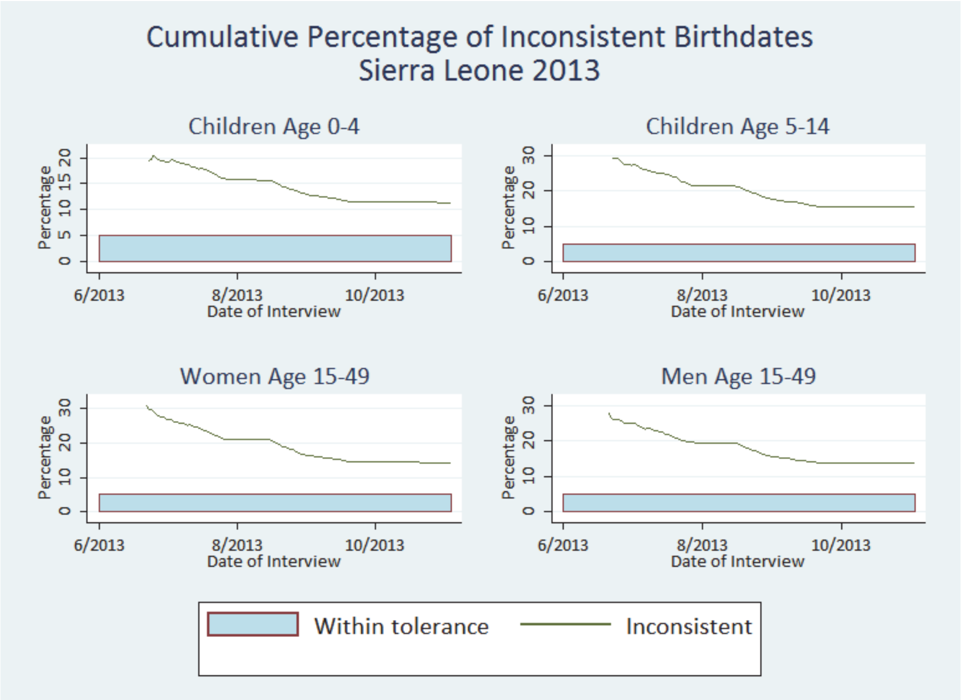
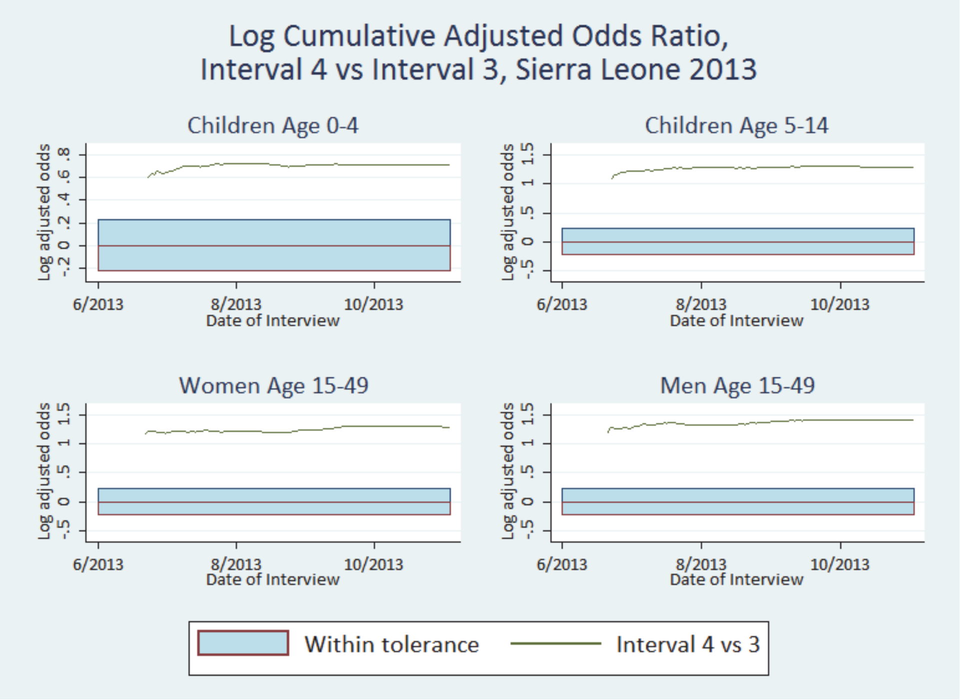


Figure 6.5 Log of the cumulative adjusted odds ratio for interval 4 versus interval 3 (YOB2) in the Sierra Leone 2013 DHS survey



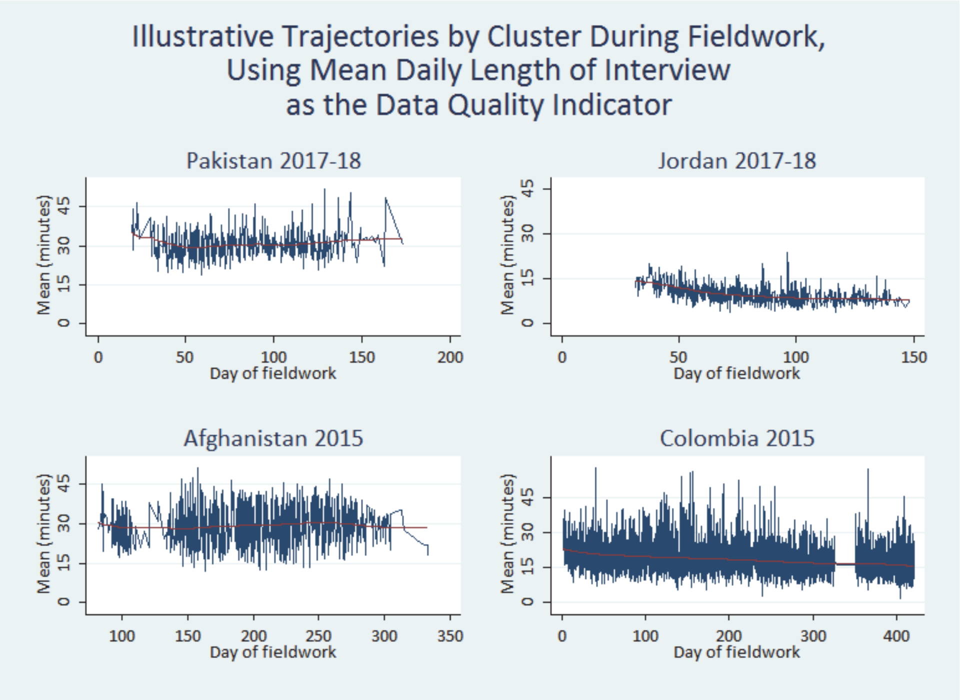
6.4 Trajectories Collapsed by Clusters and Sequenced Chronologically

Sections 6.2 and 6.3 simulated the day-to-day completion of fieldwork, first by single days and then cumulatively. Sections 6.4 and 6.5 use the clusters as units, but sequence them chronologically. In most surveys, the number of clusters is far greater than the number of days of fieldwork because many teams of fieldworkers are working simultaneously in different locations throughout the country. For that reason, the trajectories in this section are more dense and show more irregularity. Irregularities that can be traced to a smaller-than-average number of sample households in the cluster would tend to cause both upward and downward spikes.[18] To help see the difference between the approaches of sections 6.2 and 6.4, we include exactly the same four surveys in Figure 6.6 as were shown in Figure 6.1. The four subfigures of Figure 6.6 include a lowess smoother (with bandwidth 0.5) that is almost identical to the smoother in Figure 6.1.

Figure 6.6 shows very few severe downward trends, which would be problematic if they indicated that a cluster had been rushed through too quickly. The great majority of spikes tend to be upward. The surveys in Pakistan, Jordan, and even Colombia include several clusters in which the mean length of interview jumped upward. Unusually long interviews are not necessarily problematic, but a more in-depth investigation of individual surveys would examine the potential implications of such interviews.

[18] Upward and downward variation that is random would tend to be equal in magnitude on a log frequency scale, but compressed on the downward side of the frequency scale that we are using. Nevertheless, it appears visually that several of the upward spikes go beyond what we would see from random variation.

Figure 6.6 Trajectory of the length of the household interview, by cluster, sequenced chronologically during fieldwork, in the Pakistan 2017-18, Jordan 2017-18, Afghanistan 2015, and Colombia 2015 DHS surveys



6.5 Trajectories for Every Team, Collapsed by Clusters and Sequenced Chronologically

The field check tables that are used to monitor data collection are structured chronologically by interview team. We now describe how the final data files can be used to simulate the data collection in a very similar manner, following teams as they move from cluster to cluster in chronological order.

The data files include, for every case, the cluster number, the interviewer team identification number (usually the first two digits of the interviewer identification code), and the date of interview. The most recent surveys include a century day code (cdc) for the day of interview. Previous surveys include the calendar day, month, and year of interview, which can be converted to a combined code with the mdy function in Stata. If, for example, x1 or x2 are individual-level binary data quality outcomes, such as the nonresponse indicator used in Chapter 4, the following commands would be used in Stata:

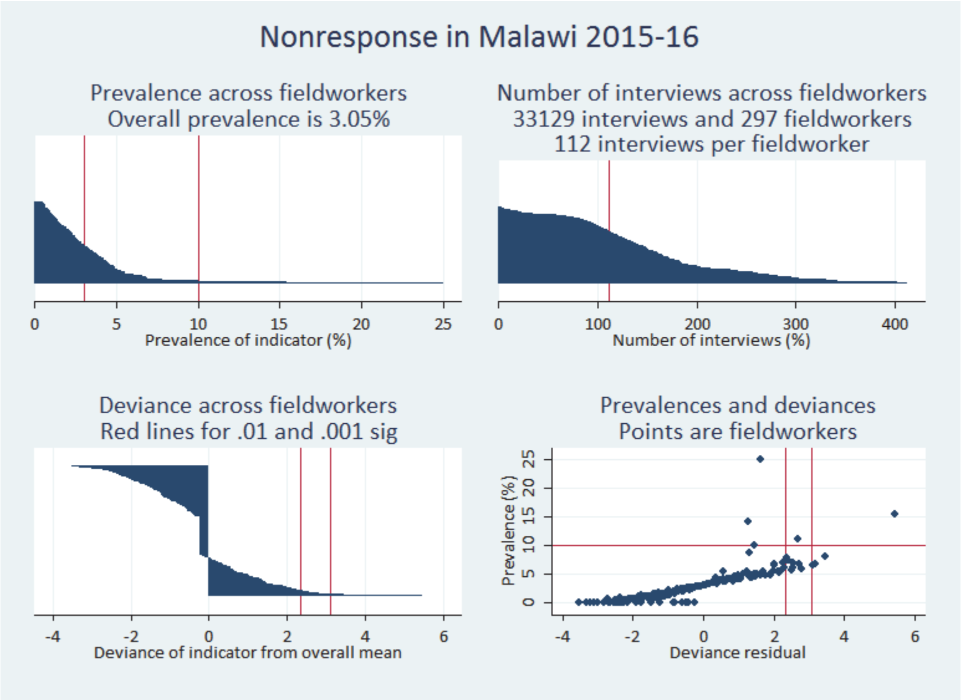
```
collapse (mean) x1 x2 date teamid, by(cluster)

sort date

gen sequence=_n
```

The example for this section is nonresponse in the Malawi 2015-16 DHS survey. Nonresponse in the individual surveys of women and men can arise from either refusal or inability to find the respondent. As emphasized in Chapter 4, DHS follows a strict policy of respecting the decisions of potential subjects to refuse to participate in the survey. Nevertheless, it is desirable to monitor the level of nonresponse. If the level is extremely low, then it is possible that the interviewers are not obtaining consent or are manipulating responses. If the level is too high, then it is possible that interviewers are not working hard enough to locate the respondents and the estimates may be biased, even after the standard adjustments to weights.

In the Malawi 2015-16 survey, the overall level of nonresponse was 3.05%, a level that most researchers would consider to be excellent. However, when variation across interviewers is analyzed, we find evidence of outliers.

Figure 6.7 includes four subfigures that lead to the identification of outliers by using a combination of their final level or prevalence of nonresponse and the statistical significance of that level. The subfigure in the upper left is a horizontal bar chart that shows the distribution of the level of nonresponse (the percentage of eligible respondents identified by these interviewers who were not subsequently given individual interviews). The vertical red line, at 3.05%, gives the mean, and another vertical red line, placed at 10%, indicates a threshold beyond which (somewhat arbitrarily) the nonresponse is unacceptably high.

Figure 6.7 Identification of problematic interviewers, using nonresponse in the Malawi 2015-16 DHS survey



In order to flag an interviewer, we apply two criteria. The first is the level or prevalence of the data quality outcome; the second is a measure of statistical significance, which takes into account the denominator, in this case the number of women and men who were eligible for an individual interview. The bar graph that is the upper-right subfigure shows the distribution of the number of interviews conducted by each interviewer. A few interviewers conducted only a handful of interviews; usually these were supervisors stepping in to assist. A few did more than 300, and sometimes even more than 400. With so much variation it is clear that the number of interviews should be taken into account.

The lower-left subfigure shows the distribution of deviance residuals around the mean level, 3.05%. Deviance residuals, described in MR24, combine the magnitude of the difference from the mean with the number of cases on which the interviewer-specific prevalence is based. They are essentially interviewer-specific z-scores for the difference from the mean. A few interviewers had large negative deviance residuals, and a few had large positive residuals. This subfigure has two vertical red lines. The one on the left identifies deviance residuals that are large enough to be significant at the .01 level; the second line indicates the .001 level. These are one-tailed p-values and are interpreted simply as thresholds or benchmarks to identify residual deviances that are too large to be attributed to random variation. The .01 and .001 one-tailed p-values are equivalent to deviances of 2.33 and 3.09, respectively.

Finally, the lower-right subfigure is a scatterplot, with a dot for each interviewer. The vertical axis shows the prevalence of the indicator (as in the upper-left subfigure), and the horizontal axis shows the deviance (as in the lower-left subfigure). The 10% threshold for prevalence is shown with a horizontal red line. The .01 and .001 thresholds for significance are shown with two horizontal red lines. Any problematic interviewers are located above and/or to the right of the red lines. A total of seven interviewers exceeded the prevalence threshold of 10% and/or the p-value threshold of .001. They are listed below:

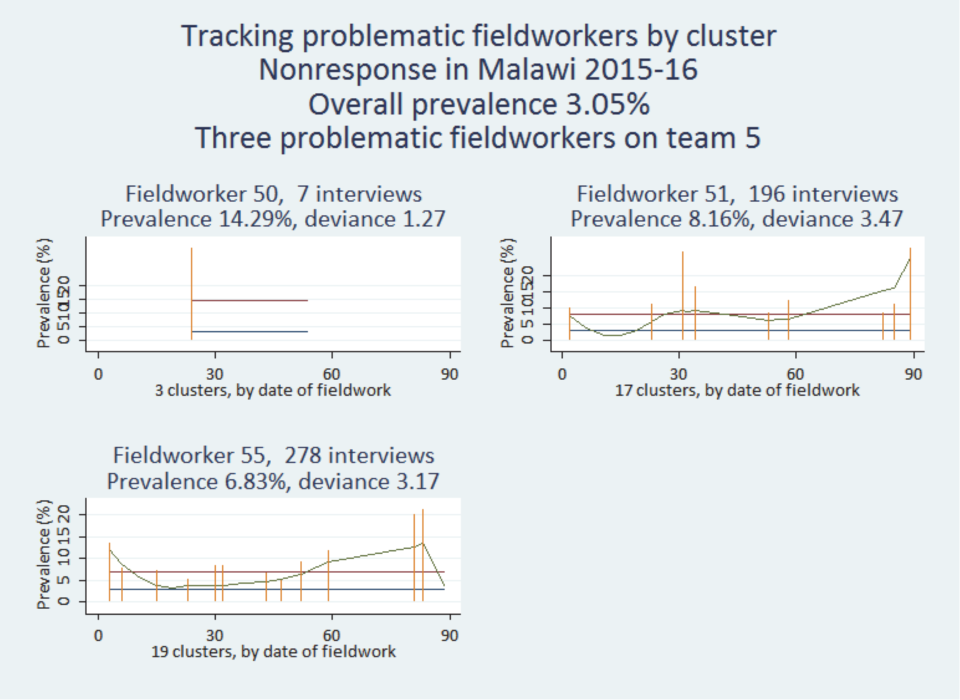
Interviewer ID	Prevalence	Cases	NR	Deviance
35	6.58%	304	20	3.12*
46	11.11%*	54	6	2.68
50	14.29%*	7	1	1.27
51	8.16%	196	16	3.47*
55	6.87%	278	19	3.17*
160	25.00%*	4	1	1.63
303	15.45%*	110	17	5.45*

The list includes the level or prevalence of the data quality indicator, as well as the denominator and numerator. The denominator (cases) is the total number of eligible women and men that the interviewer identified in all their household interviews. The numerator (NR) is the number of those eligible women and men who were not subsequently interviewed, either because of refusal or not being found at home. The prevalence is 100*NR/cases. The final column is the deviance for each interviewer.

Asterisks identify high values of prevalence or deviance. Four interviewers exceed the nominal 10% threshold for prevalence, while four exceed the nominal .001 threshold for the p-value (a deviance greater than 3.09). If nonresponse occurred at random, we would not expect even one interviewer to exceed this threshold. Although they are high, the thresholds or tolerance levels for prevalence and significance are arbitrary. It would be possible, for example, to set the prevalence threshold at 5% and the significance threshold at .05, which would greatly expand the pool of potentially problematic interviewers, but would also produce many false positives.

On the list above, a higher prevalence tended to be associated with a smaller number of cases. Interviewers 50 and 160 were team leaders who only occasionally took responsibility for an interview. Their prevalences were high but each contributed only one instance of nonresponse.

Figure 6.8 Tracking of nonresponse during fieldwork for three interviewers with flagged levels of prevalence or deviance who were on the same team (team 5), Malawi 2015-16 DHS survey

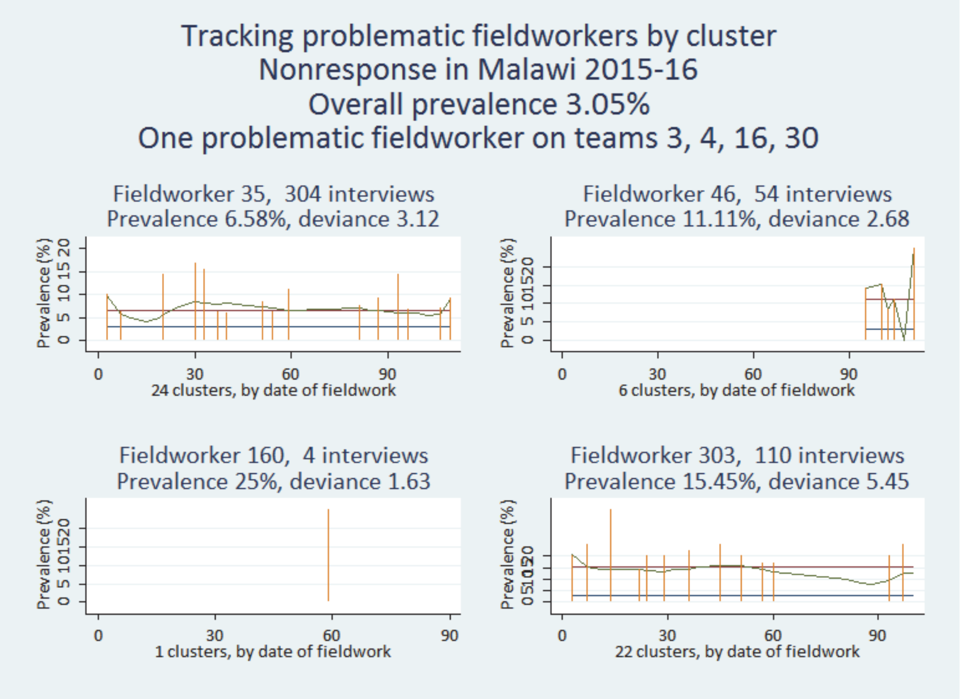


Now we turn to the main tool of this section, a simulation of tracking of the problematic interviewers during fieldwork. On the list of seven interviewers who were flagged on the basis of high prevalence and/or high deviance, three were on the same team: interviewers 50, 51, and 55 were all on team 5, and as noted, one of them was the team supervisor. Figure 6.8 shows the experience of these three fieldworkers. The subfigures, one for each interviewer, can be interpreted as follows. The horizontal axis is the day of fieldwork ("0" is the day before the beginning of fieldwork), but on a continuous scale, with each cluster assigned to the mean date of interview for that cluster. The vertical axis is the cluster-level prevalence of the outcome. Each figure has a blue horizontal straight line for the overall prevalence, which is 3.05% for this outcome, and a red horizontal straight line for the specific interviewer's prevalence, which matches the prevalences in the list. The vertical spikes identify clusters in which there was at least one instance of nonresponse. For interviewers 51 and 55, the subfigures also include an irregular horizontal line that is a partial smoothing (using lowess, with bandwidth 0.5), that smooths out the spikes and conveys more about the sequencing of nonresponse through the duration of data collection than the red horizontal straight line. Interviewers 51 and 55 had significantly higher prevalence than the overall level, but they did not exceed the 10% threshold. Both had atypically high spikes during the final days of fieldwork. The spike for the team leader, interviewer 50, occurred in a single cluster and involved a single case of nonresponse.

It may not be a coincidence that three of the seven flagged interviewers were on the same team, or that the three included the team lead. The spikes for the three interviewers did not tend to occur in the same clusters, implying that their pattern of nonresponse cannot be attributed to cluster-level characteristics. An even deeper investigation into the DQ indicators for team 5 could be justified. However, we will not go deeper here.

We also examine the other four flagged interviewers, who have id codes 35, 46, 160, and 303. Their trajectories during fieldwork are shown in Figure 6.9. Interviewer 303 is the only one on the list of seven who is flagged for both high prevalence and high deviance. Nothing more needs to be said about interviewer 160 who only did 4 interviews. Interviewer 46 only participated in a few clusters, toward the very end of the survey, and identified only 54 eligible respondents, of whom 6 were not given an individual interview. Most teams did not have a member with final digit 6, and this person may have been a late addition to the team or a substitute for another team member who had left. This is a high level of nonresponse and may be related to an unusual pattern of participation in the fieldwork. The lowess fitted lines for interviewers 35 and 303 are consistently above the horizontal lines for the overall mean prevalence of 3.05%, although both interviewers had substantial intervals in the third month—in the case of interviewer 303, the entire third month—with no refusals at all. Apart from that third-month gap, nothing stands out, such as a high level of refusals early in fieldwork, that could have drawn attention or a surge late in the fieldwork.

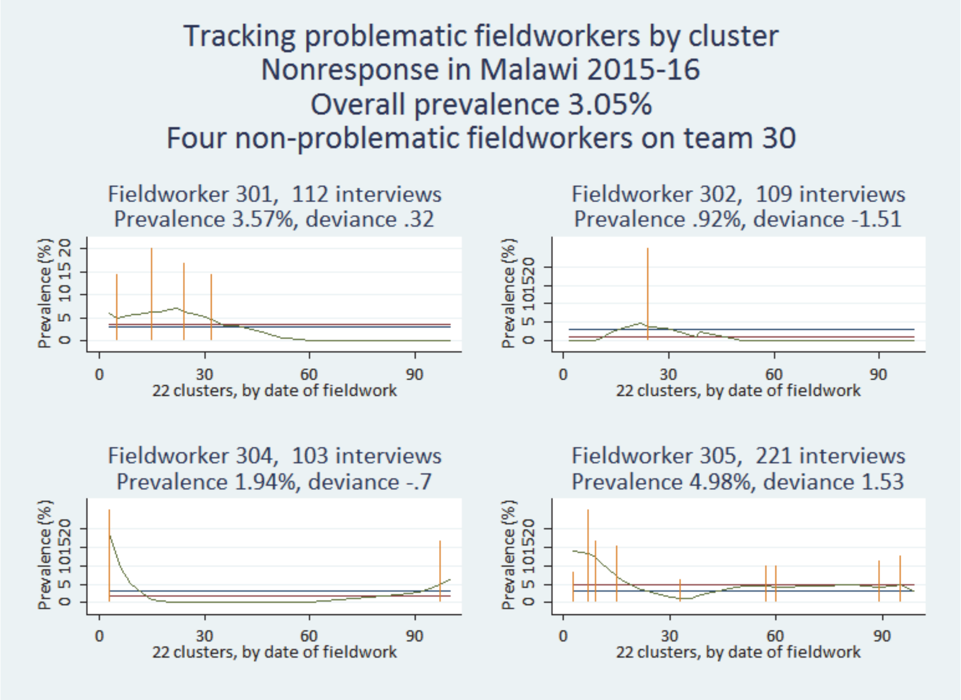
Figure 6.9 Tracking of nonresponse during fieldwork for four interviewers with flagged levels of prevalence or deviance who were on different teams, Malawi 2015-16 DHS survey



The final figure in this section, Figure 6.10, compares the most problematic interviewer, 303, with the other four interviewers on her team (301, 302, 304, and 305). These interviewers have very typical trajectories. For them, the blue and red horizontal straight lines are very close. The deviances are close to 0; two are negative, because their means were less than the overall mean prevalence of 3.05%. All of the interviewers worked in the same 22 clusters. Almost all of their nonresponse occurred during the first month of fieldwork, and it is possible that some cluster-specific factors were encountered in those early weeks. However, interviewer 303 accumulated nonresponse at a significantly higher rate than her colleagues.

Figures 6.9 and 6.10 led to the conclusion that interviewer 303 was indeed an outlier. The combination of high prevalence and high deviance for this outcome was conspicuously different from her team mates, not just from the entire pool of interviewers. Her relatively high level was not attributable to the clusters to which her team was assigned and it did not follow a pattern across the duration of fieldwork, apart from the absence of nonresponse during the third month. Nevertheless, the impact of the concentration of nonresponse that can be attributed to this interviewer was negligible. The Malawi 2015-16 household survey identified more than 33,000 eligible women and men. This interviewer was responsible, so to speak, for 110 of them, and for one reason or another, 17 of them were not interviewed. The statistical impact was certainly negligible.

Figure 6.10 Tracking of nonresponse during fieldwork for the four interviewers on team other than interviewer 303, Malawi 2015-16 DHS survey



These figures and interpretations illustrate a strategy. For the specific example, nonresponse in the Malawi 2015-16 DHS survey, nonresponse was not an issue. This was a successful survey and the level of nonresponse, from whatever cause, was quite acceptable. The goal has been to apply statistical and graphical methods to focus in on specific interviewers, dates, and clusters. The analyst could be motivated to identify specific households and individuals to identify a concentration of exceptional outcomes. It is not necessary to pursue this example, but we can summarize the steps for analyzing the trajectory of data quality during fieldwork, which could be applied to any data quality outcome:

1. Select an indicator of a potentially problematic outcome for individual cases in the data files, coded 1 if the outcome is observed, and 0 otherwise, except that it is missing or NA if it does not apply. In this example, the indicator was nonresponse.
2. Specify an upper threshold or tolerance level, above which the level or prevalence of the outcome would be considered to be serious. In this example, the threshold was set at 10%.
3. Specify a threshold or tolerance for statistical significance. The p-value refers to a null hypothesis that problematic outcomes are distributed randomly across individuals, clusters, and interviewers. Responses exceeding a high threshold are unlikely to be random. In this example, the threshold was specified in terms of a deviance residual with a p-value of .001.

4. Identify the individuals, clusters, interviewers, etc., who are above the threshold for prevalence and/or deviance. In this example, we used an “and/or” criterion, but could have moved directly to the “and” criterion, in which case we would have focused immediately on interviewer 303.
5. Check for possible contributing factors that are available in the data. It could be that other team members approached the threshold, even if not surpassing it, implying that the team was working in difficult clusters or was inadequately trained or supervised. In this example, we compared interviewer 303 with the other four interviewers on her team.

7 CONCLUSIONS

For the past 5 years, DHS has issued an average of two reports each year that are specifically concerned with the quality of DHS data. Nearly all of these reports have been actual assessments of a large number of surveys and a particular type of data, such as maternal mortality, or a particular source of error, such as interviewer effects. In general, these reports have found that DHS surveys have maintained a high standard of data quality, although they have also identified potential areas for improvement. It is important to continuously monitor the quality of the data because of the widespread use of indicators derived from DHS surveys, and the increasing use of the data files for further analysis.

Earlier reports have included substantial descriptions of methods but have been primarily data quality assessments. The present report is intended to make primarily methodological contributions, and includes actual assessments only for illustrative purposes. Most of the methods are new. The emphasis has been on frameworks or strategies for approaching the assessment of data quality, with possibilities for shifting the applications to outcomes different from those included here. For example, the approach to fertility could be applied to outcomes such as under-5 mortality, which are also derived from the retrospective birth histories, or to adult and maternal mortality, which are derived from retrospective sibling histories.

Data quality issues tend to be identified by very specific symptoms, during or soon after data collection. Many safeguards are built into the data collection process, beginning with the procedures to select, train, and supervise fieldworkers. During data collection, at frequent intervals such as every week, the survey manager examines many indicators that are tabulated by team and compared with the target values. Major deviations from the targets lead to notifications to field supervisors and, when necessary, retraining of interviewers. Many indicators fluctuate during data collection simply because of randomness. Inferring that a deviation from the target is systematic requires waiting for a sufficient number of observations.

It is often a challenge to correct an issue uncovered with the field check tables during fieldwork because of the risk of "overcorrection." For example, when ages are not known with accuracy, an effort to avoid transfers from age 15 to age 14 in the household survey can easily result in transfers in the reverse direction, from age 14 to age 15. Heaping at ages ending in 0 or 5 can easily convert to the conspicuous avoidance of 0 and 5. DHS training emphasizes the importance of data quality without identifying superficial indicators that can easily be manipulated by fieldworkers.

Alarms are raised during data collection when teams appear to spend too little time on interviews and move too quickly through the clusters. Time stamps and more accurate information about the daily movement of teams can be obtained automatically with CAPI, which greatly improves the monitoring of fieldwork.

On occasions when data quality issues are identified after data collection, it is typically the DHS staff, the implementing agency, or others who are close to the data who notice a potential issue during the preparation of the Key Indicators Report (KIR), or the main survey report. Sometimes the KIR and main report are delayed while the issue is investigated. Occasionally, a special caveat is inserted into the report about a specific indicator.

Typically, concern with a specific survey arises after fieldwork because a key indicator appears to be lower or higher than expected. For example, the estimates of recent under-5 mortality may be implausibly low. The expected level is usually based on an earlier DHS or MICS (Multiple Indicator Cluster Surveys) survey. Occasionally, an increase or a reduction was expected, and concerns are raised when there is no change from a previous estimate. For example, it is natural to question the estimates if contraceptive prevalence increased from one survey to the next while fertility was flat. When there appears to be a contradiction involving two successive surveys, it is sometimes determined that the problem was actually an under- or overestimate in the first survey, which was detected only because of an unexpected finding in the second survey.

DHS is currently conducting a thorough review and revision of the field check procedures. A companion to this report, WP162, describes new procedures for field checking of the anthropometry data for children age 0-4. The importance of biomarkers has steadily increased. New standards are being implemented, including enhanced training of the fieldworkers who measure height and weight, and new methods and criteria for monitoring data collection. More automated field checking, with dashboards to flag suspicious data, is under development.

In addition, comprehensive checks of data quality with a standard set of data quality indicators will be applied to all future surveys, as soon as the data files have been prepared. A profile for the survey will be compared with previous surveys, particularly with previous surveys in the same country. It is hoped that any weaknesses that would affect inferences about levels, trends, and differentials in the main survey indicators will be identified well in advance of the KIR. If there are problems that suggest a bias in an estimate or a relationship, then the KIR and/or the main report can mention that possibility, although, as stated earlier, it is not DHS policy to adjust the estimates.

This report has described some approaches and measures that will become part of the comprehensive assessment of all surveys. For example, the potential displacement of birthdates described in Chapter 2 and the potential disagreement between two successive surveys in their fertility rates (and similar rates) during a period of overlap before the first survey, described in Chapter 3, can be applied in an automated manner. Variation in data quality indicators according to characteristics of the respondents, as described in Chapter 4, can be assessed. If the variation is substantial, the assessment of interviewer effects, described in MR24, can use statistical models that adjust for the characteristics of respondents. For recent surveys with reliable time stamps, the relationship between the duration of the interview and other process indicators, described in Chapter 5, can be analyzed.

In addition, this report has described strategies for digging deeper into the data if potential problems are encountered during an assessment. Chapter 6 provides a link with field check approaches to data quality and illustrates how the data files can be used to identify specific clusters, interviewers, or dates with deviations from an expected value or threshold of acceptability that are both substantial in magnitude and statistically significant. The choices of threshold and level of significance can be adjusted. When there appears to be an inconsistency, it is often found to be large in magnitude but not statistically significant, perhaps because of a small denominator, or significant but small in magnitude. The alleged problem may evaporate when both criteria are invoked in combination.

Possible directions for further analysis of data quality were briefly raised in the first section of Chapter 3 and are discussed further here. Three possibilities are described with sets of questions. Methods to answer these questions can include formal models, simulation, or empirical statistical analysis.

First, how strong is the association between relatively superficial indicators of data quality, such as age heaping, and the quality of substantive indicators such as fertility and under-5 mortality? Is it possible for a survey to score badly on an age heaping measure but, in fact, to produce accurate estimates of fertility, or to score well in terms of age heaping but to produce misleading estimates of fertility? That is, do the data quality indicators actually reflect the quality of the data that are needed for decisions about policies and programs?

Second, what is the sensitivity of the substantive indicators to the data quality indicators? This question is related to the previous paragraph, but focuses on mechanical linkages. For example, how do age heaping and displacement affect the usual retrospective estimates of age-specific fertility rates and the total fertility rate? It would be possible to use a survey of established high quality, simulate different scenarios of age heaping and displacement as well as other potential influences on fertility, and recalculate the fertility rates in order to describe the sensitivity.

Third, are there dimensions to data quality? That is, are various data quality indicators associated with one another, and in what ways? For example, do heaping or rounding of various numerical responses tend to occur together? Do various types of age transfers tend to occur together? Do age heaping and age displacement tend to occur together? As another example, are respondents who omit a birth from the birth history also more likely to displace a birthdate? If the fertility data are poor, do the under-5 mortality data tend to be poor—going beyond the obvious fact that an omitted death is also an omitted birth? There is evidence of an association between incompleteness of age (and birthdate) and age heaping, but it is negative, suggesting that when interviewers are pressed to provide a number and have little solid knowledge, they (or the respondent) will tend to give a rounded value.

The most important objective is to ensure that measurement errors, or data quality problems, never reach the magnitude at which incorrect decisions are made about policies and programs. Many DHS indicators directly influence decisions about programs that improve children's nutrition and immunization status, promote maternal health, combat infectious and noncommunicable diseases, and enable couples to achieve their fertility intentions. It is essential to have data of the highest possible standard, and that requires continuously improved detection of potential data quality issues, and a better understanding of how those issues arise and can be controlled.

REFERENCES

Agarwal, N., A. Aiyar, A. Bhattacharjee, J. Cummins, C. Gunadi, D. Singhania, M. Taylor, and E. Wigton-Jones. 2017. “Month of Birth and Child Height in 40 Countries.” *Economics Letters* 157:10-13. https://econpapers.repec.org/RePEc:eee:ecolet:v:157:y:2017:i:c:p:10-13.

Ahmed, S., Q. Li, C. Scrafford, and T. W. Pullum. 2014. *An Assessment of DHS Maternal Mortality Data and Estimates*. DHS Methodological Reports No. 13. Rockville, Maryland, USA: ICF International. http://dhsprogram.com/pubs/pdf/MR13/MR13.pdf.

Allen, C., S. Namaste, T. N. Croft, and T. W. Pullum. 2019. *Evaluation of Indicators to Monitor Quality of Anthropometry Data during Fieldwork*. Working Paper No. 162. Rockville, Maryland, USA: ICF.

Assaf, S., M. T. Kothari, and T. W. Pullum. 2015. *An Assessment of the Quality of DHS Anthropometric Data, 2005-2014*. DHS Methodological Reports No. 16. Rockville, Maryland, USA: ICF International. http://dhsprogram.com/pubs/pdf/MR16/MR16.pdf.

Boerma, J. T., A. E. Sommerfelt, J. K. Van Ginneken, G. T. Bicego, K. M. Stewart, and S. O. Rutstein. 1994. *An Assessment of the Quality of Health Data in DHS-I Surveys*. DHS Methodological Reports No. 2. Calverton, Maryland, USA: Macro International. http://dhsprogram.com/pubs/pdf/MR2/MR2.pdf.

Bradley, S. E. K., W. Winfrey, and T. N. Croft. 2015. *Contraceptive Use and Perinatal Mortality in the DHS: An Assessment of the Quality and Consistency of Calendars and Histories*. DHS Methodological Reports No. 17. Rockville, Maryland, USA: ICF International. http://dhsprogram.com/pubs/pdf/MR17/MR17.pdf.

Croft, T. N., A. M. J. Marshall, C. K. Allen. 2018. Guide to DHS Statistics. Rockville, Maryland, USA: ICF.

Larsen, A. F., D. Headey, and W. A. Masters. 2019. “Misreporting Month of Birth: Diagnosis and Implications for Research on Nutrition and Early Childhood in Developing Countries.” *Demography* 56(2):707-728. https://doi.org/10.1007/s13524-018-0753-9.

MacQuarrie, K. L. D., W. Winfrey, J. Meijer-Irons, and A. R. Morse. 2018. *Consistency of Reporting of Terminated Pregnancies in DHS Calendars*. DHS Methodological Reports No. 25. Rockville, Maryland, USA: ICF. http://dhsprogram.com/pubs/pdf/MR25/MR25.pdf.

Pullum, T. W. 2006. *An Assessment of Age and Date Reporting in the DHS Surveys, 1985-2003*. DHS Methodological Reports No. 5. Calverton, Maryland, USA: Macro International. http://dhsprogram.com/pubs/pdf/MR5/MR5.pdf.

Pullum, T. W. 2008. *An Assessment of the Quality of Data on Health and Nutrition in the DHS Surveys, 1993-2003*. DHS Methodological Reports No. 6. Calverton, Maryland, USA: Macro International. http://dhsprogram.com/pubs/pdf/MR6/MR6.pdf.

Pullum, T. W., S. Assaf, and S. Staveteig. 2017. *Comparisons of DHS Estimates of Fertility and Mortality with Other Estimates*. DHS Methodological Reports No. 21. Rockville, Maryland, USA: ICF. http://dhsprogram.com/pubs/pdf/MR21/MR21.pdf.

Pullum, T. W., and S. Becker. 2014. *Evidence of Omission and Displacement in DHS Birth Histories*. DHS Methodological Reports No. 11. Rockville, Maryland, USA: ICF International. http://dhsprogram.com/pubs/pdf/MR11/MR11.pdf.

Pullum, T. W., D. K. Collison, S. Namaste, and D. Garrett. 2017. *Hemoglobin Data in DHS Surveys: Intrinsic Variation and Measurement Error*. DHS Methodological Reports No. 18. Rockville, Maryland, USA: ICF. http://dhsprogram.com/pubs/pdf/MR18/MR18.pdf.

Pullum, T. W., C. Juan, N. Khan, and S. Staveteig. 2018. *The Effect of Interviewer Characteristics on Data Quality in DHS Surveys*. DHS Methodological Reports No. 24. Rockville, Maryland, USA: ICF. http://dhsprogram.com/pubs/pdf/MR24/MR24.pdf.

Pullum, T. W., and S. Staveteig. 2017. *An Assessment of the Quality and Consistency of Age and Date Reporting in DHS Surveys, 2000-2015*. DHS Methodological Report No. 19. Rockville, Maryland, USA: ICF. http://dhsprogram.com/pubs/pdf/MR19/MR19.pdf.

Rutstein, S. O. 2018. *Data Quality Evaluation of the Niger 2017 Demographic and Health Survey*. Other Documents No. 73. Rockville, Maryland, USA: ICF. https://www.dhsprogram.com/pubs/pdf/OD73/OD73.pdf.

Rutstein, S. O., G. T. Bicego, A. K. Blanc, N. Rutenberg, F. Arnold, and J. A. Sullivan. 1990. *An Assessment of DHS-I Data Quality*. DHS Methodological Reports No. 1. Columbia, Maryland, USA: Institute for Resource Development/Macro Systems Inc. http://dhsprogram.com/pubs/pdf/MR1/MR1.pdf.

Schoumaker, B. 2014. *Quality and Consistency of DHS Fertility Estimates, 1990 to 2012*. DHS Methodological Reports No. 12. Rockville, Maryland, USA: ICF International. http://dhsprogram.com/pubs/pdf/MR12/MR12.pdf.