RAND

The Quality of Retrospective Reports in the Malaysian Family Life Survey

Megan Beckett, Julie DaVanzo, Narayan Sastry, Constantijn Panis, and Christine Peterson

DRU-2226-NICHD/NIA

December 1999

Prepared for NICHD and NIA

Labor and Population Program

Working Paper Series 99–13

The Quality of Retrospective Reports in the Malaysian Family Life Survey

Megan Beckett, Julie DaVanzo, Narayan Sastry,
Constantijn Panis, Christine Peterson*

September 1999

RAND, 1700 Main Street, Santa Monica, California 90401

* This research was supported by the National Institute on Child Health and Human Development under grant number P01 HD28372 and the National Institute on Aging under grant number T32 AG00244-03.

Abstract

In this paper, we review studies that have evaluated the quality of retrospective data collected in the 1976-77 and 1988 Malaysian Family Life Surveys. The evaluations considered the internal consistency of the data, expected relationships between variables, comparisons with external contemporaneous data sources, and analyses of replicate reports. We summarize what has been learned about data quality by subject area. The topics include: marriage; fertility and fertility-related events, such as contraception, miscarriages, birthweight, and breastfeeding; infant and child mortality; education; housing; earnings; and migration. We conclude that accuracy is greater for qualitative reports than for amounts and durations, events associated with major life changes are recalled better, and the quality of reports is greater for more educated respondents and more recent events. We discuss the implications of these findings for research and for the design of future surveys.

Preface

This is a polished version of a paper presented at the Conference on Data Quality Issues in Longitudinal Surveys at the Institute of Social Research, University of Michigan, in October 1998. This working paper is intended for users of the Malaysian Family Life Survey data as a summary of the data quality of items across various topics. There is another version of this paper that focuses in more detail on methodological issues, including telescoping and recalls error, related to retrospective data.

1. Introduction

Many social science research issues benefit from having data on events that occur over time. To reduce the cost of collecting such data, or to collect data before a baseline survey or between waves, surveys often ask retrospective questions. The desirability of collecting retrospective data is counterbalanced by important questions about its quality. For example, to what extent are past events omitted, especially those that are socially undesirable or otherwise sensitive? Does the precision of an answer decrease as the recall period lengthens? And does accuracy vary by respondent characteristics? A good understanding of the shortcomings of retrospective data is important. It may provide survey designers insights into circumstances when retrospective data will be adequate. It may also help analysts correct for—or at least anticipate—data problems, such as selective or inaccurate reports, that may otherwise yield biased or even misleading results.

In this paper, we review what has been learned about the quality of retrospective data collected in the Malaysian Family Life Surveys (MFLS). In the next section we provide a brief description of the MFLS data. This is followed, in Section 3, with an overview of the approaches to analyzing the quality of retrospective data through a review of the various techniques that scholars have taken to evaluate the accuracy of retrospectively collected information in the MFLS. Our discussion of data quality in Section 4 is organized by substantive topic so that readers can immediately focus on topic- or question-specific issues. We end the paper with a summary of what has been learned about the quality of MFLS retrospective data and discuss the implications for data analysis and survey design.

2. The First and Second Malaysian Family Life Surveys

The Malaysian Family Life Survey (MFLS) currently consists of two waves: the 1976-77 MFLS-1 and the 1988 MFLS-2.[1] MFLS-1 was conducted in three rounds, four months apart. In the second and third rounds, some questions were repeated to bring the respondents' life histories up to date. MFLS-1 interviewed 1,262 ever-married women up to age 50 and their husbands (Butz and DaVanzo, 1978; Butz et al., 1978). The MFLS-2 contained four samples; two of these, the Panel Sample and MFLS-2 New Sample, are used in the studies discussed in this paper (for more information about MFLS-2, see DaVanzo et al., 1993)[2]. The MFLS-2 Panel Sample consists of 72% of the ever-married female respondents from MFLS-1 who were successfully contacted and re-interviewed in 1988. The MFLS-2 New Sample is a sample representative of women of reproductive age (18-49), regardless of marital status, and ever-married women under age 18 living in Peninsular Malaysia in 1988. In both the Panel and New Samples, respondents' husbands were also interviewed.

MFLS-1 and MFLS-2 collected current information on many aspects of family life and full retrospective life histories on marriage, fertility and related behaviors and outcomes, child survival, work, and migration. An important feature of these data is that replicate histories are available for MFLS-2 Panel Sample respondents for the period before the 1976 MFLS-1.

3. Methods for Evaluating Data Quality

More than 250 papers have been published using the MFLS.[3] Some investigated data quality issues as their main focus; many others included a side analysis of data quality, incorporated data inaccuracy into their models, or have findings that have implications about data quality. In this section, we briefly describe the methods and identify exemplary studies—discussed in the next section in more detail—that illustrate the major methods of data quality assessment. The first two methods—comparison with external data sources and conformity to *a*

---

[1] RAND and the National Population and Family Development Board of Malaysia are planning to field the Third Malaysian Family Life Survey in 2000. It will attempt to re-interview all MFLS-1 and MFLS-2 respondents, a sample of their adult children, and a new representative sample.

[2] The two other samples in MFLS-2 were the Child Sample, consisting of adult children of women in the Panel Sample, and the Senior Sample, consisting of households with a person aged 50 or older.

[3] A list of these publications is available by contacting the Labor and Population Program at RAND.

priori expectations—are appropriate for both one-time and panel survey data; the third method, examining replicate histories, requires a panel with overlapping retrospectives.

Comparison with External Data Sources

Comparison of aggregate sample statistics to external data sources, such as government statistics and other surveys, provides an easy and useful set of data checks. Methods range from eyeballing differences in the distribution of events by age or year—for example, to determine whether there are temporal trends in quality of reported data—to using statistical tests of differences in means or rates. Consistency between the MFLS data and external sources is taken as a sign of good data quality for the MFLS (Haaga 1986; Sine 1993; Sine and Peterson 1993). However, discrepancies do not necessarily indicate misreporting problems with the MFLS data. For example, in cases where the comparison data source is another survey, discrepancies between the MFLS and comparison data may be due to differences in sample composition or question wording or biases in the comparison data set. Even where the two sources appear to be a good match, they may each be flawed. Finally, an important caveat about comparisons with external data sources concerns sample representativeness and size, which both decline with recall period. Even if the sample is representative of the population covered by vital statistics or census data at the time of the survey, it is unlikely to be representative of the population in years well before the survey because of mortality, migration, or respondents too young in the past.

External sources such as vital statistics and the censuses are generally accepted as a "gold standard" when evaluating MFLS data on fertility, infant mortality, and birthweight. Researchers have converted MFLS-1 and MFLS-2 retrospective reports on the timing of births and infant deaths into historical fertility and infant mortality rates and compared these to rates derived from vital statistics for corresponding periods (Haaga 1986; Peterson 1993; Sine and Peterson 1993). For most variables, comparison data must be obtained from other surveys. In some cases, there may be cause to believe that a comparison survey collects more accurate data. One reason is that it may have a shorter recall period. For example, Haaga (1986) compared reported contraceptive use in the 1960s and early 1970s from MFLS-1 (fielded in 1976) data on these years collected in the 1967 West Malaysia Fertility Study and the 1974 Malaysian Fertility and Family Survey. In other cases, a comparison survey may be specifically designed to elicit certain types of information and may thus collect higher quality data on these topics. In the same analysis of data quality, Haaga (1986) compared MFLS-1 rates of miscarriages, abortions, spontaneous and

induced fetal mortality with rates obtained from a survey conducted by the Federation of Family Planning Associations (FFPA) covering some of the same years (1970-73). The FFPA survey was specially designed to elicit reports of fetal mortality, incorporating several strategies absent from MFLS-1, so was considered the gold standard.

Conformity to A Priori *Expectations*

Conformity to *a priori* expectations can be assessed according to several criteria: (1) completeness of reports; (2) heaping; (3) conformity to external standards; and (4) expected relationships between or among variables. We review these each in turn.

Completeness of Reports. One aspect of data quality is the completeness of information on reported events. An important, if obvious, implication of incomplete reports is that they limit the amount of data available to the researcher and may call into question the representativeness of the sample available for analysis. We focus on the completeness of detailed follow-up questions regarding the events that were reported. For example, a respondent may report a birth but fail to provide an exact response to subsequent questions regarding, say, month and day of birth, or birthweight. In these cases, MFLS respondents were asked to provide approximate responses (e.g. approximate time of year of birth (early, middle, or late) or approximate birthweight (ranging from "very low" to "heavy").

Heaping. For many outcomes, we expect the distribution of values to be smooth. For instance, durations of post-partum amenorrhea or a child's age at death should have no peaks at 6 or 12 months, unless there are social or cultural norms. For example, when heaping on breastfeeding duration is reported, Haaga (1986) recommends looking for a culturally defined traditional weaning age as an alternative explanation.

Heaping of duration reports may be quantified by Whipple's index, defined here as six times the fraction of all duration reports that are exact multiples of six months; values above unity indicate digit preference.

Conformity to External Standards. The third approach exploits universal regularities, based on biological processes, which set a standard against which survey reports may be compared. An example of a biological standard is the sex ratio at birth (number of male births per 100 female births), which, in the absence of selective reporting or induced sex-selective abortion, should be between 104 and 106.

Expected Relationships. The fourth approach examines relationships among two or more variables in the data that are expected based on findings from other studies that can be considered valid and generalizable. We present one example. Breastfeeding has a known contraceptive effect and postpones the return of menses after a birth; the more intensive the breastfeeding, the stronger the relationship (Habicht et al. 1985). One would thus expect to find a positive relationship between the intensity and duration of breastfeeding on the one hand and the duration of post-partum amenorrhea on the other (see e.g., Habicht et al., 1985). Clinical studies can be particularly important sources from which to identify expected relationships.

Replicate Histories

Both the MFLS-1 and MFLS-2 collected full retrospective histories of marriage, fertility, birthweight, contraception, infant mortality, breastfeeding, work, and migration. For the 889 MFLS-2 Panel respondents, replicate reports are thus available for the period before the MFLS-1. The ability to match histories circumvents most of the disadvantages associated with comparisons of cross-sectional data with external sources, specifically selection effects and sample differences. A number of approaches have been taken to examine the consistency of retrospective data for reinterviewed respondents. These include comparisons of summary statistics between the two waves, comparisons of individual-level responses for the same pre-1976 events from the two waves, and comparisons of 1976 current status reports with 1988 retrospective reports on 1976 status.

The simplest evaluation of data quality using replicate histories is to compare aggregate summary statistics for the same period using the two different data sources. For instance, DaVanzo et al. (1994) compared overlapping trends in the percentage of infants ever breastfed and the percentage breastfed for 4 months or longer using MFLS-1 and MFLS-2. If the surveys yield consistent summary statistics, the trend lines will be close or overlapping for the years represented in both surveys.

Insignificant differences in summary statistics may mask inconsistency at the individual level. With replicate reports on the same events, one may assess the quality of retrospective reports for individual respondents. Several test statistics are available to evaluate individual response consistency. One such statistic appropriate for dichotomous variables is the test-retest

reliability ratio, also known as the Kappa coefficient.[4] This measure, indicative of the extent to which agreement between two reports is stronger than expected by chance, is preferable to crude percentage of agreements because it removes the effects of chance agreements (Coombs 1977). Another statistic appropriate for continuous variables is the matched pair test of the difference in means, used to identify statistically significant differences in continuous variables between samples.

Table 1 presents the Kappa coefficients for the association between current status reports for the time of the MFLS-1 survey in 1976 and 1988 retrospective reports about that point in time in 1976 for the MFLS-2 Panel sample. Current status reports should be less prone to reporting error than retrospective reports since respondents are reporting on contemporaneous events. The resulting Kappa coefficients reflect mostly recall error at the MFLS-2 interview rather than a combination of recall error at MFLS-1 and MFLS-2. The Kappa coefficients indicate low to good agreement between reports, with the highest agreement for breastfeeding status. The other Kappa coefficients indicate low levels of agreement between current status and retrospective reports on contraceptive use and work status.

Researchers have also used multivariate models to investigate factors associated with discrepancies in the replicate histories. For example, Smith and Thomas (1997) modeled the discrepancy in the reported timing of pre-1977 residential moves from MFLS-1 and MFLS-2.

4. Results from Data Quality Evaluations

In this section we summarize substantive conclusions about MFLS data quality from studies of a variety of topics covered in the survey, including marriage, fertility, contraceptive use, fetal and infant mortality, birthweight, breastfeeding and post-partum amenorrhea, education, housing characteristics, earnings, place of residence, and migration. For each topic, we first describe the questions used to elicit information in the MFLS. We then describe findings from studies that have examined the quality of the MFLS retrospective data. Despite the diversity of topics for which retrospective data quality that has been assessed, a number of consistent

[4] Defined as (O-E)/(N-E), where O is the observed number of respondents giving the same answer at two times; E is the expected number of same answers, given the marginal probabilities of responses in the two waves; and N is the number of respondents that responded both times.

correlates of data quality emerge across topics. Table 2 summarizes by topic/question the specific MFLS data sets examined in each study, the method of evaluation, main findings, and the main implications for analysis and survey design.

Marital History

Assessments of the quality of marriage histories have used data from the MFLS-2 New Sample.[5] MFLS-2 respondents were asked for their age at first marriage and number of marriages. For each marriage, beginning with first and proceeding forward in time, further questions were asked about date of marriage, date marriage ended (if applicable), and husband's occupation. If there was any discrepancy between the total number of marriages initially reported and the number of marriages described, the interviewer probed further.

Sine and Peterson (1993) examined two features of the marital history data for women in the MFLS-2 New Sample. First, they compared summary statistics from the MFLS with data from the 1970 and 1980 censuses and from the 1984/85 Malaysian Population and Fertility Survey (MPFS). Specifically, they examined the distribution of marital status by five-year age cohorts at four points in time (1970, 1980, 1985, and 1988). The MFLS contained a higher proportion of ever-married women in the 1980 and 1985 comparisons, particularly in the younger cohorts. Second, Sine and Peterson compared mean age at first marriage for women aged 25 or older who had married by age 25 by five-year age cohorts with the MPFS data. In both samples, mean age at marriage increased for successive age cohorts, as anticipated given the general trend over time toward later marriage (Sine and Peterson 1993; Sine 1993). However, consistent with the higher levels of marriage in the MFLS compared with the MPFS, MFLS-2 New Sample women reported a lower mean age at marriage than MPFS respondents.

Fertility

MFLS-1 asked women for their pregnancy histories, beginning with the first pregnancy, and moving forward through time. For each pregnancy, respondents were asked the sex, name, and date of live births; how and when the pregnancy ended and the duration of pregnancy for non-live births; number of months pregnant for currently pregnant women; antenatal care for all pregnancies; further information about live births (prematurity, place of delivery, whether still

alive, infant feeding, etc.); and timing of onset of menstruation after each pregnancy outcome. In MFLS-2, in addition to date of birth, respondents were asked their age at time of the birth. MFLS-2 also added an introductory set of "Brass" questions that asked women about the total number of pregnancies and living and deceased children. These questions appeared to improve recall of births in MFLS-2 as evidenced by the greater number of pre-1977 births newly reported in MFLS-2 relative to the number apparently forgotten between MFLS-1 and MFLS-2 (Peterson 1993).

Sine (1993) and Sine and Peterson (1993) considered the proportion of reported births with missing date of birth information in the MFLS-2 New Sample. Date of birth is necessary for estimating fertility and infant mortality rates. In the MFLS-2 New Sample, approximately six percent of birth dates were reported without the month and two percent without month and year, about the same levels of incomplete birth dates as was found in "other recent retrospective surveys in Asia" (Sine 1993; Sine and Peterson, 1993). There was poorer reporting as one moves backwards in time, which may lead to increasingly biased estimates of fertility and infant mortality rates for earlier periods. Nevertheless, MFLS-2 did remarkably well in eliciting mother's age at birth when mothers could not provide the child's year of birth: 100% completeness (Sine 1993). With this piece of information, year of birth can be imputed, which is useful for some types of analysis (e.g., trends in fertility rates). However, it is not sufficient for analyses that require more exact date of birth information (e.g., investigations of the time between births and subsequent conceptions).

Haaga (1986) examined the multivariate correlates of live births with missing or inexact birth dates.[6] Less educated women and ethnic Malays and Indians were least likely to provide complete or exact birth dates.

Investigators have examined the completeness of the MFLS fertility history in three additional ways. First, they examined the sex ratio at birth for evidence of births being omitted according to the child's sex. The expected value of the sex ratio is 104 to 106 live male births for every 100 live female births. MFLS-1 sex ratios indicated no systematic under-reporting of female births (Haaga 1986). Similarly, there were no significant differences in sex ratios between

[5] MFLS-1 is of limited value for this purpose because it was restricted to ever-married women.

[6] Instead of providing precise dates, some respondents responded with time of year (e.g. early, middle, or late in the year).

the MFLS-2 New Sample and Malaysian vital statistics (Sine and Peterson 1993). Second, researchers compared age-specific marital fertility rates implied by retrospective histories with external data sources. Haaga (1986) calculated age-specific marital fertility rates for 1956-60 and 1961-65 using MFLS-1 and compared these with rates based on Malaysian vital statistics from the beginning and end of each 5-year interval. In almost all cases, the MFLS-1 retrospective reports were bounded by the vital statistics rates. The mean number of children ever born to married women in the MFLS-2 paralleled—though were slightly lower than—the rates implied by the 1970 and 1980 vital statistics (Sine and Peterson 1993). Finally, Haaga (1986) looked for evidence of event displacement, as described by Potter (1977), whereby births that occurred more than five years before the interview data were reported as occurring more recently than they really did. In the case of event displacement, births should cluster around the cut-off point; beyond the cut-off point, fertility is accurately reported. Haaga (1986) found no evidence of such clustering.

To summarize, quality assessments based on sex ratios, children ever born, and age-specific marital fertility rates do not indicate systematic omission or misreporting of births in the MFLS-1 or in the MFLS-2 New Sample. The only evident problems in these data are that exact birth dates were not always reported and omission of birth dates was higher for births that occurred further back in time. This omission may influence analyses that depend on specific timing of births. Analyses that only need crude timing of birth, such as estimating age- or year-specific fertility rates, can rely on imputed year of birth.

Contraceptive Use

For each interval between pregnancies, as well as the interval between marriage and the first pregnancy, MFLS-1 respondents were asked whether they had used any means to avoid pregnancy. Two probes were used. First, for periods of more than two years during which the respondent was not pregnant or contracepting, interviewers were instructed to probe for missed pregnancies or periods of abstinence and spousal separation. Second, if a respondent did not identify breastfeeding as a form of contraception in periods during which she breastfed, interviewers were instructed to ask about the concurrent use of other contraceptive methods. If there was a period of more than three years during which the respondent was neither pregnant nor contracepting, interviewers were instructed to probe for missed pregnancies or periods of spousal separation or abstinence from intercourse.

Haaga (1986) was concerned that the definitions of traditional contraceptive methods and specific probes used in MFLS could affect comparisons with other studies. For example, MFLS-1 included breastfeeding as a traditional method whereas other studies did not. Haaga (1986) compared percentages of women in 10-year age cohorts in the MFLS sample who were married and reported ever use of modern contraceptives or sterilization before 1967 with similar estimates from the 1967 West Malaysia Fertility Study and the 1974 Malaysian Fertility and Family Survey. The MFLS-1 percentages are lower than the 1967 and 1974 comparison data sets, with similar sized differences in both the 1967 and 1974 comparisons. In other words, the problem is not so much with misreporting the timing of events in the more distant past as with general under-reporting in the use of modern contraceptives. Haaga hypothesized that this problem was due to less probing in MFLS-1, which had more topics to cover besides fertility.

Sine and Peterson (1993) compared three measures of contraceptive use for the MFLS-2 New Sample with data from the MPFS (1984/85). First, they compared the proportion of ever-married women who had used contraceptives prior to January 1985. Overall, the prevalence of ever use in the two surveys is similar (74% for the MFLS-2 New Sample; 77% for the MPFS). Next, they compared the contraceptive methods used by women reporting current use in January 1985. The MFLS-2 New Sample and MPFS compared closely on methods, although there was a higher share of "efficient methods" in the MFLS (64% for the MFLS-2 New Sample; 58% for the MPFS), with most of the discrepancy due to the pill. Two explanations can account for these differences. One is that the MFLS-2 figure is based on retrospective reports about 1985, whereas the MPFS figure is based on current status at the time of that survey. Respondents may have been more likely to recall methods that worked than "inefficient" methods, which are on average less effective. The other explanation is that the MFLS-2 New Sample was somewhat younger than the MPFS sample (the oldest MFLS-2 New Sample woman was only 46 in 1985 compared to age 49 in the MPFS) and younger women may have preferred the pill. Finally, Sine and Peterson examined the distribution of current methods among MFLS-2 New Sample women who said they were contracepting at the time of the 1988 interview. Looking at current use reduces problems associated with selective recall of efficient methods and age composition effects. The 1985 and 1988 distributions for the MFLS-2 New Sample were very similar. For example, 64% reported use of an efficient method in 1985 and 1988. Given that there should have been little change over three years in the distribution of methods, this suggests that selective recall and age

composition differences do not account for the differences between the MFLS-2 and MPFS samples in 1985.

Overall, retrospective reports of contraceptive use and the distribution of methods approximate those reported in other surveys; there is no evidence that the MFLS, especially the MFLS-2 New Sample, suffered from systematic bias relative to other surveys.

Infant and Child Mortality

In MFLS-1 respondents were asked, for each child born alive, where the child was at the time of the survey. If child had died, respondents were asked for the date of death. If this could not be recalled, the respondent was asked age of child in weeks at time of death to provide sufficient information to calculate and record date of death.. Two modifications were employed in MFLS-2. Brass questions were used to prompt respondents' memory of all of their births. The rationale is that questions about individual pregnancies (and their outcome and survival status) would better prompt for deaths than a question about total number of deaths. Second, MFLS-2 asked for the age at death rather than the date of death. This second change created some challenges for comparing reports of infant/child mortality across the two surveys (Peterson 1993). In the Panel Sample, exact date of death (day, month, year) was provided in MFLS-1 for only 30% of cases, while 45% of the matched deaths had just month and year, and 25% had only year. Thus, for 75% of the matched deaths only an approximate age at death could be calculated from MFLS-1 data.

Several researchers have employed a variety of approaches to assess the quality of infant mortality data in the MFLS-1, MFLS-2 New Sample, and the Panel Sample. Overall, retrospective reports of infant mortality were of high quality. Haaga (1986) showed that annual infant mortality rates (IMR) implied by MFLS-1 data were consistent with annual IMRs obtained from vital statistics between 1945 and 1974 (with corrections for the different age distribution in the MFLS-1 in the earlier years). Sine and Peterson (1993) replicated this comparison using MFLS-2 New Sample data. They also concluded that IMR trends in the MFLS-2 New Sample tracked the vital statistics IMRs, with both sources showing substantial declines from 1968-1986. However, MFLS-2 data tended to have *higher* levels of infant mortality than vital statistics data. This finding runs contrary to the assumption that retrospective data would tend to result in under-reporting of sensitive topics, such as infant mortality, relative to vital statistics. Sine (1993) examined reporting of infant mortality from another angle. He compared sex ratios at death of

children in Malaysian vital statistics and MFLS-2 data. Vital statistics rates indicated a very consistent sex ratio at death of 138 to 140 deaths of male children under age five for every 100 deaths of females in the same age range across three birth cohorts (1960's, 1970's, and 1980's). Sex ratios across these three birth cohorts in the MFLS-2 data were also consistent and, with one exception, did not differ (statistically) significantly from vital statistics rates. Moreover, this exception was the opposite from the expected difference. In the 1970's birth cohort, males appeared to be *underreported* relative to female births. Again, no evidence of under-reporting of female births or deaths is evident in the MFLS data.

Although of generally high quality in terms of reporting that a death occurred, information on timing of or age at death was somewhat problematic. On the one hand, the MFLS-1 survival curves for the first year of life and MFLS-2 distributions of reported age at death for the babies who died compared closely to those obtained from vital statistics (Haaga 1986, Sine and Peterson, 1993). However, the MFLS-2 age-at-death distribution indicated modest levels of heaping at six-month intervals. Although the heaping at six months of age does not influence infant mortality rates (deaths within the first 12 months of life), heaping at 12 months does. It is impossible to determine from a reported age at death of exactly 12 months whether the death actually occurred before or after one year of age.

Peterson (1993) explored the analytic implications of using reported age at death to construct categorical variables indicating that the child died within a certain age interval or period. She examined the level of agreement or reliability of reported age at death for births that were reported by the matched sample in MFLS-1 and MFLS-2. In the aggregate, agreement[7] of two constructed variables (whether child died by 1976 and whether child died in first year of life) was 99%. At the individual level, agreement was also almost perfect (99%) for the variable indicating whether the child died in first 12 months of life. Successively finer categories, however, yielded diminishing agreement, with 90% agreement for whether the child died in first six months of life, and 71% for seven finer categories of age at child's death (0-3, 4-6, 7-12, 13-24, 25-36, 37-60, and 61+ months).

[7] Defined as the proportion of births for which the same report on survival status was provided in MFLS-1 and MFLS-2.

Fetal Mortality

In MFLS-1, respondents were asked to report all pregnancies, whether they ended in a live or in a non-live birth. For intervals of two or more years between pregnancies during which contraception was not used, interviewers asked whether any additional pregnancies had occurred, including any that did not result in a live birth. The length in months of each pregnancy that ended without a live birth was ascertained. No attempt was made to distinguish between induced and spontaneous abortions although respondents were asked why they thought they lost the baby. In MFLS-2, respondents were asked if anyone "did anything to end this pregnancy" (if pregnancy ended without a live births before seven months) and about the outcome of each pregnancy (if pregnancy ended without a live birth after seven months). These questions were used to determine if the termination was due to a spontaneous abortion, induced abortion, or miscarriage.

Under-reporting of fetal mortality was a serious problem. Haaga (1986) compared MFLS-1 rates of spontaneous and induced fetal mortality with rates obtained from a survey conducted by the Federation of Family Planning Associations (FFPA) covering some of the same years (1970-73). This survey was designed to elicit reports of fetal mortality, incorporating several strategies to do so. For example, women were first asked if they had had any unusually delayed menstrual periods, then if they thought they might be pregnant, and finally whether they had done anything to induce menstruation. Relative to the MFLS, the FFPA should serve as a "gold standard" for the level of fetal mortality. The overall FFPA fetal mortality rate (92 per 1000 pregnancies) for the time period 1970-73 was considerably higher than the MFLS-1 rate (53 per 1000 pregnancies) for the same time period. This suggests that the MFLS-1 suffered from substantial underreporting of fetal mortality, although it is impossible to deduce if induced abortions were underreported to a greater extent than spontaneous abortions. Sine and Peterson (1993) replicated this comparison of fetal mortality rates using the MFLS-2 New Sample. MFLS-2 New Sample fetal mortality rates were about 10 points higher than in MFLS-1, suggesting that the additional probes did elicit some additional reports, but were still considerably lower than FFPS fetal mortality rates, again leading to the conclusion that the MFLS data probably under-estimated fetal mortality for the period 1970-73.

Since MFLS-2 collected information on spontaneous abortions and induced abortions separately, it was possible to investigate which type of pregnancy loss was more likely to be underreported. The percentage of MFLS-2 New Sample women who reported ever having experienced a spontaneous abortion was almost identical to that found in the 1984/85 Malaysian

Population and Fertility Survey (15% for MFLS-2 New Sample; 16% for MPFS). However, rates of induced abortion reported by the MFLS-2 New Sample were half that of the MPFS data (3% for MFLS-2 New Sample; 6% for MPFS). It is important to note that MFLS-2 asked the question on induced abortion in the same way as MPFS so that differences in question wording cannot explain the different rates. As a final data quality check on internal consistency, Sine and Peterson examined the age trend in the proportion of women ever experiencing fetal mortality by 1985. This proportion should rise with the age of the woman. Among MFLS-2 women, the anticipated pattern was found: a greater proportion of women in successively older cohorts had ever lost at least one fetus.

Although the age trends implied by retrospective reports of miscarriage in MFLS-2 appear reasonable, time trends do not. Over a period characterized by improving health conditions and declining infant, child, and maternal mortality, data from the MFLS-2 New Sample imply *rising* rates of miscarriage (Panis and Lillard 1994). From roughly 1950 to 1985, rates of fetal mortality more than doubled from 5% of all pregnancies initiated in a given year to more than 10%.[8] Panis and Lillard posit that women forgot about miscarriages over time. Miscarriage in the first three months of pregnancy is mostly biological, unassociated with behavioral factors. Panis and Lillard documented that increases in rates of miscarriage over time were due to first-term miscarriages, and, moreover that the most educated women were more likely to report miscarriages. In other words, miscarriages that occurred long ago and miscarriages that occurred to less educated women were probably under-reported. They further hypothesized that given the very low rate of reported induced abortions in the MFLS (Malay and Chinese women reported 53 abortions for 6,333 pregnancies; Indian women were dropped from the analysis because of too few cases), some women may have misclassified induced abortions as spontaneous abortions or miscarriages, particularly for more recent pregnancies.

Birthweight

MFLS-1 and MFLS-2 both collected information on birthweight in the same manner. Women were first asked to report the exact birthweight of each child. Exact birthweight, in this

[8] One source of variation, undoubtedly, is change in the age composition of the New Sample from 1950 to 1985: the oldest woman in the New Sample would have been 11 in 1950 and 46 years old in 1985. This example highlights the necessity of ensuring that cohorts or samples being compared have comparable age compositions.

case, means that number of pounds was reported, but not necessarily ounces. Although women were asked to report in pounds and ounces, 12% of women who reported pounds and ounces reported zero ounces (i.e., there was heaping on zero ounces) (Sine and Peterson 1993). Given 16 ounces to a pound and assuming a random distribution of birthweight, six percent of birthweights would be expected to have zero ounces.

If a woman did not know the exact birthweight, she was then asked to provide an approximate birthweight using five categories, ranging from "very low" to "heavy". Exact birthweights were reported for 70% of MFLS-1 births (DaVanzo et al. 1984) and for 90% of MFLS-2 New Sample births (Sine and Peterson 1993). In the aggregate, a constant proportion of the matched sample provided exact birthweights for pre-1977 births (33.0% in MFLS-1; 32.6% in MFLS-2) (Peterson 1993).

Generally, birthweight information in the MFLS was of good quality. The frequency distributions of exact birthweights for the MFLS-1 sample of live births between 1970 and 1976 was comparable to the frequency distribution of birthweights based on 1977 vital statistics (DaVanzo et al. 1984). The distribution of exact birthweights for births between 1982 and 1986 in the MFLS-2 New Sample compared well to the 1984 vital statistics distribution (Sine and Peterson 1993). The reliability of exact birthweights for births in the matched sample was good. In the aggregate, mean birthweight showed a small but statistically significant decline between MFLS-1 (6.8 lbs.) and MFLS-2 (6.7 lbs.). Individual-level agreement paralleled that of the aggregate comparisons. For two-thirds of the births, the two reports were within a half pound. However, when they differed, there was a tendency toward lower birthweights in MFLS-2. This was most likely due to rounding down to the nearest pound (e.g., a 7lb., 4 oz. baby in 1976 becomes 7 lbs. in 1988). Among matched births, two thirds of birthweights reported in 1976 for births in 1976 or earlier had pounds and ounces reported while only 50% of 1988 reports about those same births had reports on non-zero ounces.

Another way to consider birthweight is to examine whether the weight is above or below some meaningful biological threshold (Peterson 1993). For example, low birthweight, commonly defined as at or below 5.5 lbs., is associated with a significantly increased risk of mortality and morbidity and may indicate prematurity or inadequate fetal nourishment. Among pre-1976 births reported in both MFLS-1 and MFLS-2, the difference in the proportion of low birthweight births in the two surveys was small and nonsignificant (11.3% in MFLS-1 compared to 12.6% in

MFLS-2). Similarly, in the matched sample the 91% of responses agreed about low-birthweight status in the two surveys (reliability statistics showed that this agreement was not due to chance).

MFLS asked for approximate birthweight values because of the nontrivial proportion of cases in which respondents were unable to report an exact birthweight. Omission of these birthweight values would have introduced potentially serious sample selection bias because the births with exact birthweights differed from those with inexact reports. For example, a multivariate analysis concluded that exact birth weights were more likely to be reported for births that took place in a hospital or clinic, more recent births, those where the child survived at least that year, and those to mothers who were highly educated, resided in urban area, and was Chinese or Indian (DaVanzo et al. 1984). Failure to include births without exact birthweights could produce spurious associations between birthweight and its correlates.

The validity of approximate birthweight responses has been examined in two ways. First, DaVanzo et al. (1984) replicated the comparison of the distribution of vital statistics birthweights using the full MFLS-1 sample (exact birthweights plus approximate birthweights) by assigning values for approximate birthweights (based on the mean and standard deviations of the exact birthweight data). As was the case with the exact-birthweight-only MFLS-1 sample, the distributions for the full MFLS and vital statistics were comparable. Second, approximate birthweights appeared to be valid measures. In both MFLS-1 and MFLS-2 New Sample, "very low" and "low" birthweights were significant predictors of infant mortality (Haaga 1986; Sine and Peterson 1993), consistent with the association between low exact birthweight (≤5.5 lbs.) and infant mortality (DaVanzo et al. 1984). Among respondents who provided an exact birthweight in MFLS-1 but only a size response in MFLS-2, the size response in MFLS-2 was "consistent with MFLS-1 birthweight" (Peterson 1993). Although longer recall periods decrease the ability of women to recall exact birthweights, women seem to have been able to recall the relative size quite well over time. However, the reverse was not true. Women who initially provided an approximate birthweight in MFLS-1 followed by an exact birthweight in MFLS-2 were not more likely to provide consistent reports. Peterson (1993) also found some evidence of regression towards the mean among births from the matched sample for which a size response was provided at both waves: 64.8% of birthweights were "average" in MFLS-1 compared with 76% in MFLS-2. The analytical impact of this tendency would be a dilution of the estimated effects of birthweight (e.g., on infant mortality) for reported births with longer recall.

Further evidence regarding the quality of retrospective reports of birthweight, based on exact and approximate birthweights, comes from comparing analyses based on these data with clinical studies. DaVanzo et al. (1984) concluded that the relationships between birthweight and the child's sex, birth order, and survival status, and the mother's age and socioeconomic status that emerged from the MFLS-1 data are the same as those found in clinical studies that employed clinically measured birthweights.

Breastfeeding and Post-Partum Amenorrhea

MFLS respondents were asked, for each child, whether they breastfed the child and, if so, the duration of breastfeeding and the age at which supplementary food was first introduced. MFLS respondents were also asked how long it was until menstruation resumed after each pregnancy. For both breastfeeding and amenorrhea, answers were coded according to the time units used by the respondents -- days, weeks, months and years. At Rounds 2 and 3 of the MFLS-1, women were asked again whether their first child had been breastfed, and if so, for how long. At Round 2, women were asked to name the first type of milk other than breast milk given to their first-born and most recently born children. In Round 3, women were asked a more general question, this time for the first-born, second-born, and most recently born children, about the first food or liquid given to the child on a regular basis. Because of concerns about heaping, if a duration was reported as "about one year" or "about six months," interviewers were instructed to probe for whether breastfeeding was less or more than that duration.

There was a slight change in the main breastfeeding question between MFLS-1 and MFLS-2. In MFLS-1, the question "Did you breastfeed (NAME OF CHILD)?" was followed by the qualifier "I want to know even if it was just for a few days," whereas in MFLS-2 the qualifier was changed to "I want to know even if you just tried once or twice." This change was made in an effort to identify more unsuccessful attempts to breastfeed (DaVanzo et al. 1994). Unfortunately, it has interfered with attempts to evaluate whether trends in MFLS data reflect behavioral change.

The quality of retrospective reports on breastfeeding in MFLS has been studied extensively. Data quality was highest for the "did/did not" questions and less good for "how long" questions (Peterson 1993). The test-retest reliability ratio between MFLS-1 Rounds 1 and 2 (4 months apart) for whether respondent ever breastfed her first child is 0.91 (Haaga 1986); the Kappa coefficient (0.70) for ever-breastfed among women in the matched MFLS-1/MFLS-2

sample (on reports 12 years apart) indicated good agreement (DaVanzo et al. 1994). The prevalence of women who ever breastfed their youngest child, as of 1985, in the MFLS-2 New Sample was broadly comparable, though higher, than in the MPFS (80% in MFLS-2; 77% in MPFS) (Sine and Peterson 1993).

The proportion of women ever breastfeeding in MFLS-2 was higher than in MFLS-1. DaVanzo et al. (1994) plotted the percentage of Malaysian infants breastfed in MFLS-1 and in the MFLS-2 New Sample by year of birth of infant, from 1956 to 1985. The same upward movement from MFLS-1 in the percentage of infants who were breastfed around 1975 (Haaga 1986) is also apparent in the MFLS-2 New Sample data, although the MFLS-1 rates are lower than those from MFLS-2 for the overlapping years. The same tendency for higher levels of ever breastfeeding was seen in the matched sample of pre-1977 births as well. Although the majority of Panel respondents provided the same response about whether they breastfed in both surveys, on average they were significant more likely in 1988 to have reported breastfeeding for pre-1977 births (87.2%) than they were to have reported this for these same births in 1976 (84.3%) (DaVanzo et al. 1994). Women were more likely to change their 1976 report from "not breastfed" to "breastfed" in 1988 than they were to do the reverse. One-third of the non-breastfed infants in MFLS-1 were subsequently reported as breastfed; in contrast, only three percent of breastfed infants in MFLS-1 were later reported as not breastfed (Peterson 1993). DaVanzo et al. (1994) concluded that the increased breastfeeding rate in MFLS-2 probably reflected two factors: first, changes in the way MFLS qualified the main breastfeeding question; second, women may have underreported breastfeeding when it was less popular (the mid-1970s) and/or overreported when it was more popular (the late-1980s).

The quality of data on breastfeeding duration is more problematic than data on whether or not the child was breastfed. Short-term consistency of reported duration of breastfeeding was good: the kappa coefficient for responses to the question "Did you breastfeed your first child?" in rounds 1 and 2 was .91 (Haaga 1988). However, consistency between breastfeeding durations reported in both MFLS-1 and MFLS-2 was poor. Although the difference in the aggregate mean duration calculated for the Panel sample respondents was small, it was significant (12.7 months in 1977; 12.1 months in 1988) (Peterson 1993). Reported duration of breastfeeding showed considerable variability between MFLS-1 and MFLS-2 in the matched sample of births. Nearly two-thirds of durations differed by more than a month and nearly half differed by more than three months, with more recent births showing the largest differences (Haaga 1988; Peterson 1993).

Another result of memory decay seemed to be a shift to reporting in less precise units of time: 60% of births where duration was reported in days or weeks in MFLS-1 were reported in months in MFLS-2; 25% of births with duration reported in months in MFLS-1 were reported in years in MFLS-2 (Peterson 1993).

Breastfeeding duration data are seriously susceptible to heaping on responses that are multiples of six months (Haaga 1988; Klerman 1995). Haaga (1986) computed Whipple's index, a summary measure of digit preference, for reported breastfeeding durations by year of birth using the MFLS-1 data. Whipple's index was 2.94 for 1946-55 births, 2.23 for 1955-64 births, and 1.78 for 1965-74 births, indicating that digit preference was more common for births that occurred further in the past. (Recall an index of 1=no digit preference.) Mothers of Malay descent, with less education, and reporting on a birth in the more distant past had a higher probability of heaping (Haaga 1988). Failure to account for such measurement error may cause a large bias (usually towards null results) in models estimating correlates of breastfeeding duration or effects of breastfeeding duration on infant health outcomes (Haaga 1988; Klerman 1995).

Heaping need not always be indicative of poor data quality. For example, if health professionals recommend that women breastfeed for 12 months or there is a norm to wean babies around their first birthday, one would expect to find a corresponding peak in reported breastfeeding durations. Haaga (1986) examined this issue using two approaches. First, he looked at the anthropological literature to see if there were known norms in Malaysia with respect to duration of breastfeeding; he did not find any such evidence. Second, he compared the distribution of retrospective responses to current status reports. If heaping reflects socially regulated behavioral—rather than reporting patterns—the same distribution of responses seen in retrospective data should be evident in contemporaneous reports. Haaga, however, found no evidence of heaped responses in the MFLS-1 current status reports, and concluded that the heaping of retrospectively reported duration of breastfeeding was probably not portraying actual behaviors. Similarly, Haaga found heaping for post-partum amenorrhea, where there was no biological reason to expect it.

Peterson (1993) considered whether constructing meaningful categories of breastfeeding duration improved the reliability of these measures in the matched sample of births. The World Health Organization recommends children be breastfed for four months without supplementation (DaVanzo et al., 1994). A larger proportion of the MFLS-2 New Sample reported breastfeeding durations greater than four months compared to MFLS-1 reports for the same overlapping years,

particularly for more recent births. However, comparison of the MFLS-2 New Sample and the MFLS-1 sample can be influenced by differences in age compositions. A more definitive test is to compare reports for the matched sample of infants. In the aggregate, there was no significant difference between 1976 and 1988 prevalence of breastfed greater than four months in the matched sample (Peterson 1993); at the individual level, the Kappa coefficient (0.68) indicates good reliability (DaVanzo et al. 1994). In other words, collapsing duration of breastfeeding into dichotomous variable improved the test-retest reliability of this measure. Finally, the MFLS-1 did a better job of eliciting short-term breastfeeding durations than the 1974 Malaysia World Fertility Survey WMFS (which asked women how many months they breastfed, if they breastfed). Despite issues of reliability and heaping of breastfeeding duration data in the MFLS, reported duration of breastfeeding measure behaved as expected in empirical analyses.

Despite the concern about the validity of the retrospective reports of breastfeeding duration in the MFLS data, reported durations correlate as expected with other biological processes. Habicht et al. (1985) and VanLandingham (1993) demonstrated that the relationship between reported breastfeeding duration in MFLS-1 and post partum anovulation and between breastfeeding duration and waiting time until next conception was as expected based on many other studies across several different countries.

Kuate Defo and DaVanzo (1996) evaluated the reliability of reported reasons for no or short breastfeeding in the matched sample. In the aggregate, there was remarkable similarity in the two distributions. The ranked order of importance for various reasons and whether they were more important as reasons for no vs. short breastfeeding was nearly the same in both surveys. The most important reason for not initiating breastfeeding or for short breastfeeding in both MFLS waves was no milk or insufficient milk, consistent with other studies. In cases where reasons change between surveys, these changes were not substantively important. For example, the reason with greatest variability across waves was "other"; only 12% of women who said "other" in MFLS-1 did so in MFLS-2. The second greatest source of variability involved women who reported maternal or child illness as a reason in MFLS-1; most of the response changes, however, were to other health-related categories. Reliability was greatest for reasons linked to specific events, such as mother returning to work or child dying, further suggesting these data were informative for linking events and behavior.

Haaga (1988) examined the reliability of reported type of first supplementary food at Rounds 2 and 3 of MFLS-1. Agreement for infants whose first supplementary food or liquid was

another type of milk was good ($\kappa^2 = .73$). Length of recall significantly decreased the reliability of reported first food (except for Chinese mothers), as did rural residence. Overall, this analysis indicated that women were able to consistently recall, in the short-run, whether the first supplementary food or liquid was another type of milk.

Post-partum amenorrhea is the period after pregnancy before regular menses returns and its duration is positively related to the duration and intensity of breastfeeding. Like with breastfeeding, in MFLS-1 there was substantial heaping at six-month intervals (Haaga 1988). Moreover, the correlates of heaping on post-partum amenorrhea were similar to the correlates of providing a preferred digit response for duration of breastfeeding. In multivariate regression, longer recall periods increased the probability of providing a heaped response (indicating memory decay) as did Malay ethnicity and less education. A chi-square test of two variables, one indicating a peak response for duration of breastfeeding and the other indicating peak response for duration of amenorrhea, showed that digit preference for one variable was associated with digital preference for the other (Haaga 1986). In other words, there was a propensity for respondents to express digit-preference across questions.

The designers of MFLS anticipated digit-preference for breastfeeding and amenorrhea (Haaga 1986). Interviewers were instructed to probe if a duration reported as "about one year" or "about six months" was more likely to have been less than or greater than the approximate duration, in an effort to decrease heaped responses. However, it is not possible to determine from the data whether this probe was successful in reducing heaping.

Education

There were two questions in the MFLS surveys concerning the number of years of formal schooling and whether the respondent obtained a certificate. Additionally, the life history portion of the MFLS-1 questionnaire also asked whether respondents attended a school, college, or university at any time after their 15th birthday. If they did, respondents were then asked more detailed questions about schooling in late adolescence, including whether they were enrolled in school at age 15. In the second round of the MFLS-1, respondents were asked directly whether they were in school at age 15. These data were used to examine the internal consistency of school attendance questions (Haaga 1986). The reliability ratio for the Round 2 response to whether or not the respondent was in school at age 15 compared with the answer implied by the life history (Round 1) was very low: 0.11 (based on a sample of 1,155 women). Of the 109 women in Round

2 who said that they had been in school at age 15, almost all (107) failed to mention being in school at age 15 in their life histories (Round 1). It is probable that respondents were confused by the initial screening question, which asked about schooling "at any time after your 15th birthday." Respondents may have interpreted this to mean to at age 16 rather than at age 15.

Housing Characteristics

As part of the migration history, women were asked for all housing moves since marriage or age 15, whichever occurred first. They were also asked about the characteristics of each house in which they resided, including whether it had electricity, piped water and the type of toilet. Haaga (1986) compared the reported presence of basic household amenities in 1970 from MFLS-1 with data from the 1970 housing census data. The MFLS-1 data compared well with data from the census. In MFLS-1, 50 percent of households reported having electricity, compared with 44% of households reporting electric lighting in the census. Sixteen percent of MFLS-1 households reported a flush toilet and 24 percent reported no toilet, compared to 19 percent and 20 percent, respectively, from the census. Finally, 48% of households in both MFLS-1 and the census reported piped water in 1970. Tong (1993) provides indirect evidence that the MFLS-2 reported levels of housing amenities is consistent with 1970 and 1980 census data. Trends in housing standards, as measured by source of water supply, type of toilet facilities, and electricity have been steadily improving when data from the 1970 and 1980 censuses is combined with current status reports from the MFLS-2.

Earnings

In MFLS-1 and MFLS-2, female respondents and their husbands were asked for a complete history of earnings since age 15 or first marriage, whichever occurred first. In both surveys, they reported their income at the beginning and end of each job. In MFLS-1 alone, work and earnings information was collected at three-year intervals for all jobs held more than three years. We should note that MFLS-1 male respondents were even more select than the sample of female respondents. In particular, the sample included men who were currently married to a woman under age 50.

Smith (1983) modeled MFLS-1 retrospective reports of monthly earnings by husbands for all years beginning in 1949. The estimated deflated income growth rate of 2.4% per year was virtually the same as the growth in Malaysian Gross Domestic Product of 2.3% per year between

1950 and 1973. This is a close correspondence considering the selectivity of men in the MFLS data. Smith's specification allowed for persistent unobserved individual effects that were uncorrelated across individuals. Although Smith pointed out that retrospective wage data probably have large variance because of memory decay, he did not model this form of heterogeneity. Lillard and Kilburn (1995) did. They showed that the variance in reporting error associated with earnings rose with the length of the recall period.

Place of Residence and Migration

Migration histories were collected in the same way in both MFLS-1 and MFLS-2. Each female respondent and her spouse gave their location (district and state) at birth and at age 15. They were then asked about each location where they lived for at least three months since age 15. Women were asked to report every time they changed house, while men were asked about moves across districts. Moves from Malaysia to another country were counted as migration events; moves between foreign countries were not.

Assessments of the MFLS migration histories have focused on the matched sample. Smith and Thomas (1997) examined the correspondence between location at birth and at age 15 reported in MFLS-1 and 12 years later in MFLS-2 for men and women. Over 95% of respondents placed themselves in the same state at birth and at age 15 at both times. Respondents had a little more difficulty placing themselves in the same district at birth and at age 15 at both waves, with about 85% so doing. When the area considered is expanded to include adjacent districts—to account for districts that have split into two or more districts at a later point in time—the percentage of respondents who place themselves in the same district or adjacent district increases to over 90%.

Smith and Thomas (1997) also examined the reliability of migration histories, covering all male inter-district moves and female inter- and intra-district moves between age 15 and the time of the MFLS-1 interview. For each MFLS-1 inter-district move (and each intra-district move for women), a search was made across all MFLS-2 inter-district moves to identify one with which the MFLS-1 move could be matched. A non-match occurred if a move reported in MFLS-1 was not reported in MFLS-2 or if the definition of a district boundary changed between surveys. About 67% of intra-district moves for women reported in MFLS-1 were also reported in MFSL-2, and about 80% of inter-district moves for men and women reported in MFLS-1 were reported 12 years later. In other words, longer distance moves were better recalled.

For moves that were reported in MFLS-1 and MFLS-2, it is possible to measure the reliability of timing for successfully matched moves. Smith and Thomas (1997) concluded that there was a slight tendency for dates reported in MFLS-2 to be more recent than those reported in MFLS-1—i.e., migration events were telescoped.

Smith and Thomas (1997) estimated the respondent and event characteristics associated with reporting a move in MFLS-2, conditional on the move being reported in MFLS-1. They also examined the correlates of discrepancy of timing of reported moves in both surveys. Models were estimated for men and women separately. Better responses were provided by women compared to men and by younger and more educated respondents. On the whole, moves associated with more salient events were reported more reliably. Consistency was higher and discrepancy in reported timing lower for moves that were: long-distance, made jointly by spouses, noncircular (i.e., the respondent did not return to place of origin after a short interval), coincident with other salient life events, crossed a district boundary (for women), or occurred closer to the time of the interview. Also, fewer total number of moves decreased the discrepancy of reported timing for women, but not for men. Interestingly, there were sex differences in types of coincident life events that improved recall. For women, marriage and birth of a child improved reliability of reported moves. For men, moves associated with job changes were better recalled. Finally, for men, the discrepancy in timing was smaller if an interviewer judged the responses as highly reliable. In other words, interviewer assessments of (men's) responses were consistent with the data quality tests examined.

5. Implications for Analysis and Data Collection

We now summarize data quality patterns that emerge across topics and discuss their implications for data analysis and data collection.

Data quality deteriorates with the length of the recall period. This finding emerged from studies that used the MFLS cross-sectional and panel data to analyze a range of topics. Topics ranged from birth dates and breastfeeding duration to migration events. These studies employed a variety of techniques for assessing data quality, ranging from comparisons with external data sources to analysis of correlates of reliable reporting across waves. Birth dates were more completely reported and heaping on reported duration of breastfeeding was less severe for more recent births (Sine and Peterson 1993; Haaga 1986). Lillard and Kilburn (1995) found that residual variation of earnings rose with the length of the recall period. Discrepancies in the

timing of moves between MFLS-1 and MFLS-2 reports were smaller for more recent moves (Smith and Thomas 1997); and test-retest reliability (the likelihood that a move reported in 1977 was also reported in 1988) was higher for more recent moves (Smith and Thomas, 1997).

Socioeconomic characteristics predict quality of retrospective reports. Less educated persons, ethnic Malays, and rural residents had lower test-retest reliability of moves (Smith and Thomas, 1997), increased heaping in breastfeeding duration and post-partum amenorrhea (Haaga 1986), and increased use of approximate instead of an exact birthweight reports (DaVanzo et al., 1984).

Events that are socially undesirable or unacceptable are more poorly reported. This finding is consistent with a large literature that has examined the social acceptability of events and the willingness and ability of respondents to provide accurate reports about these events (Sudman and Bradburn 1974). Abortion—particularly induced abortion—appeared to be underreported in MFLS-1 and MFLS-2 (Haaga 1988; Sine and Peterson 1993; Panis and Lillard 1994), compared to contemporaneous surveys in Malaysia that focused on fertility and reproductive health. It may have also been the case that reports on less socially sensitive behaviors, such as breastfeeding, were also responsive to social norms. Reported breastfeeding rates for pre-1976 births were higher in MFLS-2 than in MFLS-1 (DaVanzo et al. 1994). One hypothesis is that changes in social norms regarding breastfeeding may have induced women to overreport previous breastfeeding behavior when it was popular (in 1988) or to underreport previous breastfeeding when it was less popular (in 1976).

Events that are more salient to respondents are more accurately reported. Moves were more consistently reported if they involved longer distances, were non-circular, crossed an administrative boundary, or were coincident with other salient events, like a marriage or new job (Smith and Thomas 1997). Respondents were more likely to report an exact rather than approximate birthweight if the child survived its first year (DaVanzo et al. 1984) and were more likely to report a reason for no or short breastfeeding if it was tied to a salient event (e.g., death of a child) (Kaute Defo and Davanzo, 1996).

Accuracy is greater for qualitative outcomes than for amounts and durations. Peterson (1993) showed that both in the aggregate and at the individual level, responses to yes/no questions (such as, "Have you ever breastfed?") were reported more reliably than responses to questions about how much and how long (e.g., "How long did you breastfeed?"). Even when durations and amounts were reported, she found that the correspondence between reports was

better when the response was collapsed into categories (e.g., birthweight above or below 5.5 pounds, duration of breastfeeding more or less than four months).

Addressing Data Quality in Descriptive Statistics and Analyses

The consequences of recall error for data analysis depends on its nature and on whether the report is an outcome or explanatory variable. As indicated above, recall error may take the form of underreporting of events, mistiming of events, heaping, or other mismeasurement of duration or amounts. Explanatory variables affected by recall error will result in coefficient estimates that are biased towards zero, a standard result for the effects of measurement error. Klerman (1995) confirmed this finding in an analysis of the effects of breastfeeding duration on infant health outcomes. Note that if the errors in explanatory variables are correlated with other explanatory variables, however, coefficients may be biased in either direction.

For variables that represent outcomes of interest, recall error may result in many types of biases. Underreporting of live births when the child subsequently died will result in infant mortality estimates being biased downward. Mistiming of an event, such as a birth or death, can affect period fertility and infant mortality rates, trends in fertility and mortality rates, and estimates of lengths of birth intervals. Heaping at 12 months for the age at which a child died may bias estimates of infant mortality rates upward (since deaths occurring say, at 13 months, are reported as having occurred at 12 months). Finally, in multivariate analyses, measurement error amounts to reduced precision of the outcome—i.e., greater residual variation or possibly heteroskedasticity.

One method for incorporating measurement error in data analysis or descriptive statistics is to create categorical responses from continuous variables. Peterson (1993) showed that this improved the consistency of reports on breastfeeding and birthweight. Constructing categorical variables from continuous reports works best when the loss of information is minimal. This may occur when there is a biological basis for the transformation—for example, converting continuous birthweight into an indicator of low birthweight and converting age at death into an indicator of infant mortality (age at death $\leq$12 months). However, one should use caution when there is heaping at the cutoff point.

Analysts can also incorporate measurement error directly into multivariate models. Variables that are associated with recall error should be incorporated into models. Their absence may represent an omitted variable problem, which could result in systematic biases in estimates

of all parameters. Analysts can also incorporate heteroskedasticity into multivariate models to obtain more efficient estimates. Lillard and Kilburn (1995) explicitly specified the variance of the error term to be a function of recall period, respondent education, and interviewer's assessment of the participation effort of the respondent. They found that the variance of wage reports increased with recall period and decreased with both respondent education and interviewer's assessment of the respondent's understanding of the questions. Heaping leads to reduced precision in reported durations. Panis and Lillard (1994, 1995) addressed this problem in a hazard model analysis of infant and child survival with an approach that allowed them to smooth out duration effects at certain important intervals, such as one year.[9]

Implications for Data Collection

The findings of our review of studies examining data quality in the MFLS lead to some suggestions for survey design that may help mitigate recall error when data is collected through retrospective reports.

Provide benchmark information. In panel surveys, respondents' recall may be improved by reminding them of their situation at the time of the previous wave and bringing them forward from that point. This also reduces interview time, relative to administration of a complete retrospective questionnaire. A caveat may apply if a question involves some concept, such as a health condition, about which the general public becomes more educated over time. For example, a respondent may have indicated lack of hypertension in the baseline, only to realize later that he or she did in fact suffer from the condition at that time but did not know its proper term. The survey designer has to allow for outcomes where this could be the case. Also, benchmarking is not costless; the interviewer has to be provided with easy access to previous responses when reinterviewing a respondent. A final issue is that many retrospective histories begin in the present

[9] When asked about the age at which a child died, MFLS-1 and -2 let respondents specify the time unit (day, week, month, year). This enabled Panis and Lillard (1994, 1995) to reduce the consequences of heaping by adjusting the precision with which they constructed survival outcomes. They developed a discrete hazard model in which the likelihood of a death was equal to the difference in survivor values at the beginning and end of the window during which the death occurred. For example, the likelihood of a death that occurred after "twelve months" equaled the difference in survivor values at 11.5 and 12.5 months, whereas the likelihood of a death after "one year" was equal to the difference in survivor values at 0.5 and 1.5 years. The precision of the outcome measure was thus equal to the time unit that the respondent specified.

and work back in time; however, for benchmarking to work best, the retrospective history should begin at the time of the last interview and work forwards.

Ask alternative questions if the respondent is unable to provide a precise response. Data quality of events in the more distant past may be improved by asking alternative questions when respondents are unable to provide a response. For example, if a person does not remember an exact date, he or she may be able to provide the age at the time of the event or the time of the year (early, middle, late) it occurred. Other examples of follow-up questions in the MFLS included approximate birthweight if an exact birthweight could not be reported. In all cases, the survey designer needs to confirm that the distributions of the approximate responses are consistent with that of the exact responses or with external data sources (see e.g. DaVanzo et al. 1984 for approximate birthweights reports).

Use a calendar. Responses may be improved by using a calendar on which major national events (e.g. key holidays, national events, social unrest, or natural disasters) are pre-printed, and on which salient life events are recorded. In particular, this may improve internal consistency and sequencing of events. For example, did the move take place before or after the marriage? Did the discontinuation of contraception occur before or after the death of a child? Did the respondent re-enter the workforce before or after the onset of the economic crisis? Use of a calendar has been shown to improve data quality (Becker and Sosa 1992).

Train interviewers to [prompt and] probe. A well-trained interviewer may be able to elicit information on underreported events. MFLS interviewers, for example, were instructed to inquire about a miscarriage, induced abortion, missed episodes of contraceptive use, or spousal separation if the period between two births was more than three years. For reported duration or amounts that may be subject to heaping at critical values, probing may help correct classification. For example, MFLS interviewers were instructed to probe if respondent reported "about one year" or "about six months" for whether the duration was more likely to have been shorter or longer than the reported value. If the calendar method is not used, interviewers should be trained to cross-reference events in different life areas (as the MFLS did).

First ask summary questions. Details on sensitive areas may be better retrieved if the questions are preceded by one or more summary questions. For example, in MFLS-2 Brass questions inquire about the total number of pregnancies and living and deceased children. The interviewer then followed up with detailed questions about outcomes for each pregnancy and prompted the respondent if the totals did not match. The MFLS-1 did not ask Brass questions

although it did ask "How many children have you had?" Peterson (1993) found newly reported births in the MFLS-2 for the period prior to the MFLS-1 interview date, which, she concludes, may have resulted from the use of the Brass questions.

Vary recall period with saliency. More salient life events are remembered better, as Smith and Thomas (1997) have shown for migration. Thus, one may restrict the recall period for common, non-salient events. Radloff (1983) makes this recommendation for collecting information on short-distance, return moves when collecting a migration history. He recommended that short-distance and short-term migration be collected for the preceding twelve months only, while more salient moves, such as non-circular moves or those involving crossing an administrative boundary, be collected for longer recall periods. A less extreme recommendation is that the robustness of results based on short-distance and short-term migration events reported for a long interval should be compared to results based on the subset of the most recent such moves.

Ask detailed questions about a subset of events of a particular type. Anecdotal evidence from MFLS interviewers suggests that women who had many pregnancies anticipated questions related to third and higher-order pregnancies and answered "the same as the previous child" to many questions about topics such as birthweight and place of delivery. It is not clear whether this tendency reflects intra-woman correlation across births, respondent fatigue, interviewer fatigue, or is a function of decreasing saliency to respondents of higher-order births (and by extension, other events). One approach to improving data quality on repeated events is to limit the number of repeated events that are asked about. For example, use a random subsample or select a subsample for which the respondent is likely to have the best recall or which is of the most interest to the analyst. However, this approach may not always work. The Encuesta Guatemalteca de Salud Familiar collected detailed information about recent illnesses experienced by a woman's two youngest children (Peterson et al. 1997). Interviewers reported similar problems, where the woman would give information on the youngest child and then would break-off during the questioning for the 2nd child or would respond "the same, the same." Future work is needed both to document the extent to which reported information on specific events declines with higher order events and ways to address the problem.

Be sensitive to respondent's culture, language, and other characteristics. Questions may be better interpreted and responses more freely given if interviewers are carefully matched to respondents. For example, MFLS field supervisors attempted to assign individual interviewers to

respondents of the same gender, ethnicity, and native language. At this time, we can only speculate that these efforts improved data quality; in the future, we will empirically examine this issue.

References

Becker, Stan, and Doris Sosa (1992), "An Experiment Using a Month-by-Month Calendar in a Family Planning Survey in Costa Rica," *Studies in Family Planning,* Vol. 23, No. 6, pp. 386-91.

Butz, William P., and Julie DaVanzo (1978), *The Malaysian Family Life Survey: Summary Report,* RAND, R-2351-AID

Butz, William P., and Julie DaVanzo, Dorothy Z. Fernandez, Robert Jones, and Nyle Spoelstra (1978), *The Malaysian Family Life Survey: Appendix A, Questionnaires and Interviewer Instructions*, RAND, R-2351/1-AID.

Coombs, Lolagene C. (1977), "Levels of Reliability in Fertility Survey Data," *Studies in Family Planning*, Vol. 8, No. 9, pp. 218-32.

DaVanzo, Julie, and William P. Butz, (1978), *The Malaysian Family Life Survey: Summary Report.* RAND, R-2351-AID.

DaVanzo, Julie, John G. Haaga, Tey Nai Peng, Ellen H. Starbird, and Christine E. Peterson (1993), *The Second Malaysian Family Life Survey: Survey Instruments*. RAND, MR-107-NICHD/NIA.

DaVanzo, Julie, Jean-Pierre Habicht, and William P. Butz (1984), "Assessing Socioeconomic Correlates of Birthweight in Peninsular Malaysia: Ethnic Differences and Changes Over Time," *Social Science and Medicine*, Vol. 18, No. 5, pp. 387–404.

DaVanzo, Julie, Jeffrey Sine, Christine Peterson, and John Haaga (1994), "Reversal of the Decline in Breastfeeding in Peninsular Malaysia? Ethnic and Educational Differentials and Data Quality Issues," *Social Biology*, Vol. 41, No. 1-2, pp. 61-77.

Fleiss J. L. (1981), *Statistical Methods for Rates and Proportions (2^{nd} ed.)*. Wiley, New York.

Haaga, John G. (1986), *The Accuracy of Retrospective Data from the Malaysian Family Life Survey*. RAND, N-2157-AID.

Haaga, John G. (1988), "Reliability of Retrospective Survey Data on Infant Feeding," *Demography*, Vol. 25, No. 2, pp. 307-14.

Habicht, J. -P., Julie DaVanzo, W. P. Butz, and Linda Meyers (1985), "The Contraceptive Role of Breastfeeding," *Population Studies*, Vol. 39, pp. 213-32.

Klerman, Jacob A. (1995), "Insights into Heaping from Retrospective Breastfeeding Data," in *Proceedings of the International Conference on Survey Measurement and Process Quality,* Bristol, England, April 1995, pp. 239-44.

Kuate Defo, Barthelemy, and Julie DaVanzo (1996), "Data on Reasons for No or Short Breastfeeding: Are They Reliable and Do They Help Us Understand Infant Feeding Behavior?" RAND, DRU-1305-NICHD.

Lillard, Lee A., and M. Rebecca Kilburn (1995), "Intergenerational Earnings Links: Sons and Daughters," RAND, DRU-1125-NIA.

Panis, Constantijn W. A., and Lee A. Lillard (1994), "Health Inputs and Child Mortality: Malaysia," *Journal of Health Economics*, Vol. 13: 455-89

Panis, Constantijn W. A., and Lee A. Lillard (1995), "Child Mortality in Malaysia: Explaining Ethnic Differences and the Recent Decline." *Population Studies*, Vol. 49, No. 3, pp. 463-79.

Peterson, Christine E. (1993), "Long-Term Recall of Fertility-Related Events: The MFLS Experience." Paper presented at the 1993 meeting of the Population Association of America, Cincinnati, OH.

Peterson, Christine E. (1997), *The 1995 Guatemalan Survey of Family Health (EGSF): Overview and Codebook*, RAND, DRU-1538/3-NICHD.

Potter, J. E. (1977), "Problems in Using Birth-History Analysis to Estimate Trends in Fertility," *Population Studies,* Vol. 31, No. 2, pp. 335-64.

Radloff, Scott R. (1983), "Detecting Migration: An Exploration of Measurement Issues Using the Malaysian Family Life Survey," RAND, N-1927-AID.

Sine, Jeffrey (1993), "Data Quality in Retrospective Surveys," in *Proceedings of the Seminar on the Second Malaysian Family Life Survey, Kuala Lumpur, Malaysia, October 1991*, RAND, CF-109-NICHD/NIA/WFHF, pp. 96-112.

Sine, Jeffrey, and Christine E. Peterson (1993), *The Second Malaysian Family Life Survey: Quality of Retrospective Data for the New Sample*. RAND, MR-110-NICHD.

Smith, James P. (1983), "Income and Growth in Malaysia," R-2941-AID, RAND.

Smith, James P., and Duncan Thomas (1997), "Migration in Retrospect: Remembrances of Things Past," DRU-1628-NICHD.

Sudman, Seymour, and Norman M. Bradburn (1974), *Response Effects in Surveys: A Review and Synthesis*, Chicago, IL: Aldine Publishing Company.

Tong, Foo Sya (1993), "Basic Amenities in Household in Peninsular Malaysia," in *Proceedings of the Seminar on the Second Malaysian Family Life Survey, Kuala Lumpur, Malaysia, October 1991*, RAND, CF-109-NICHD/NIA/WFHF, pp. 162-70.

VanLandingham, Mark (1993), "Breastfeeding and Waiting Time to Conception for Malay Women: A Tale of Two Surveys," *Social Biology*, Vol. 40, No. 3-4, pp. 215-23.

Table 1. Kappa Coefficient for the Comparison of 1976 Current Status Reports with MFLS-2 Retrospective Report on 1976 Status, Panel Sample (n=889)

1976 status	K^2	Level of agreement[a]
Marital status		
Currently married?	.43	Good
Married, separated, divorced, widowed	.42	Good
Contraceptive use		
Practicing any form contraception?	.38	Poor
Type of contraception (or none) (17 categories)	.37	Poor
Work status		
Ever worked?	.39	Poor
Work status (6 categories)	.39	Poor
Breastfeeding status		
No, partial, or complete breastfeeding	.75	Good

[a] Values of K^2 greater than 0.75 reflect excellent agreement beyond chance, between 0.40 and 0.75 indicate good agreement beyond chance, whereas values below 0.40 represent poor agreement (Fleiss 1981).

Table 2: Summary of studies evaluating quality of retrospective data in MFLS

Topic		MFLS data set and sample	Method of evaluation	Findings	Implications	*
Marriage		MFLS-2 New	Comparison with 1984/5 Malaysian Population and Fertility Survey (MPFS).	Single women are under represented because institutional settings not sampled in MFLS-2. Rapid industrialization in 1980s may explain fewer single women living at home in MFLS-2 (1988) than in 1984/5 MPFS.		12
		MFLS-2 New	Conformity to external standard and comparison with 1984/5 Malaysian Population and Fertility Survey (MPFS) and Malaysian Census.	Mean age at first marriage is rising for successive cohorts in MFLS-2 New Sample and MPFS, as anticipated. New Sample reports lower mean level of marriage across all cohorts, compared to MPFS and the census.		11
Fertility	Birth dates	MFLS-2 New	Completeness of reports, etc.	About 6% births reported without month, 2% without month and year. 100% of births had mother's age at time of birth.	Collect mother's age at time of birth in order to impute year of birth date.	12
				Trend toward less complete birth reporting as proceed backwards in time.	Use calendar method to link birth dates with other salient events (understanding that these other events may also be reported inaccurately). Be cautious about quality of data for earlier periods.	11
		MFLS-1	Completeness of reports.	10% of birth dates were missing exact month of birth or reported only by season. Ethnic Malays and Indians were more likely to be missing month of birth or to provide an inexact birth date (time of year).	Be more sensitive to cultural and communication differences between interviewer and respondent in interviewer training.	3
	Sex ratio	MFLS-1	Conformity to external standard.	Sex ratios at birth are between 104 and 106, indicating no systematic under reporting of female births.		3
		MFLS-2 New	Comparison with vital statistics.	Sex ratios do not statistically differ from sex ratios based on vital statistics.		12

Topic		MFLS data set and sample	Method of evaluation	Findings	Implications	*
	Children ever born	MFLS-2 New	Comparison with other survey data 1984/85 Malaysia Population and Fertility Survey (MPFS).	No statistically significant differences in the number of children born per ever married woman New Sample and MFFS.		12
	Marital age-specific fertility rates	MFLS-1	Comparison with vital statistics.	Marital age-specific fertility rates (ASFR) are within range of the rates based on Vital statistics data. No evidence of event displacement.		3
		MFLS-2 New	Comparison with vital statistics.	Marital ASFRs across periods parallel, but are generally lower than, vital statistics rates.		12
Contraceptive use		MFLS-1	Comparison with 1967 West Malyaisa Fertility Study and 1974 Malaysian Fertility and Family Survey.	Compared with other surveys, MLFS-1 shows lower prevalence of modern contraceptive use or sterilization at all ages and length of recall periods.	Probe in more detail about contraceptive use (to obtain fuller information.	3
		MFLS-2 New	Comparison with 1984/85 Malaysian Population and Fertility Survey (MPFS)	Among users, a higher proportion report an 'efficient method'. May be due to forgetting of inefficient methods.	Probe for inefficient methods.	12
Child and infant mortality	Infant mortality	MFLS-1	Comparison with vital statistics.	Infant mortality rates in the MFLS sample show the same downward trend from 1945 to 1975 as Malaysian vital statistics.		3
		MFLS-2 New	Comparison with vital statistics.	MFLS-2 New and vital statistics infant mortality rates (IMRs) from 1968-1986 track each other closely.		12
	Sex ratio at death of children	MFLS-2 New	Comparison with vital statistics.	No statistically significant differences between sex ratios at death based on MFLS-2 and vital statistics (138 to 140 male child deaths for every 100 female deaths).		11

Topic		MFLS data set and sample	Method of evaluation	Findings	Implications	*
	Infant's age at death	MFLS-1	Conformity to external standard.	An anticipated, MFLS-1 infant survival curves show that most deaths occur very soon after birth and there is a long right-hand tail.		3
		MFLS-2 New	Conformity to external standard and with vital statistics.	Distribution of age at death such that most deaths occur very soon after birth and there is a long right-hand tail. Compared with vital statistics, MFLS-2 has higher levels of infant mortality in earlier years.	May be that registering infant deaths which occurred in rural Malaysia was more difficult 20 years ago than it was for women to recall those deaths 20 years later.	11
				There is heaping of deaths at six and 12 months.	Heaping at 12 months biases in unknown direction estimates of infant mortality rates.	
	Whether child died by 1976	Matched sample	Matched aggregate comparison and 1:1 match	In the aggregate, no statistical differences between the MFLS-1 and MFLS-2 proportions of children reported to have died by 1976. In 1:1 match, 99% agreement on whether child died.	In survey design, use "yes/no" questions in addition to eliciting duration or timing information. If have duration or timing information, for analysis construct discrete meaningful categories from such data.	10
	Child died in first 12 months of life	Matched sample	Matched aggregate comparison and 1:1 match	In the aggregate, no statistical differences in the MFLS-1 and MFLS-2 proportions of children reported to have died in first 12 months of life. In 1:1 match, 99% agreement on whether child died in first 12 months of life.	In survey design, use "yes/no" questions in additio to eliciting duration or timing information. If have duration or timing information, for analysis construct discrete meaningful categories from such data.	10
	Child died in first 6 months of life	Matched sample	1:1 match	90% agreement for whether infant died in first six months of life.	In survey design, use "yes/no" questions in addition to eliciting duration or timing information. If have duration or timing information, for analysis construct discrete meaningful categories from such data.	10
	Age child died (7 categories)	Matched sample	1:1 match	Level of agreement is 71%. Reporting complicated by fact that date of death reported in MFLS-1 and age at death in MFLS-2. Poorer level of agreement is obtained with narrower categories.	In analyses, construct as broad of categories as feasible.	10
Fetal mortality		MFLS-1	Comparison with Federation of Family Planning Associates survey (FFPA, 1970-73).	Overall, MFLS fetal mortality rates are about 40% lower than the FFPA survey. The FFPA survey was especially designed to elicit information on fetal mortality.	Use more detailed methods to elicit information on highly sensitive topics (see text for example from the FFPA survey to collect information on fetal mortality).	3

Topic		MFLS data set and sample	Method of evaluation	Findings	Implications	*
		MFLS-2 New	Comparison with Federation fo Family Planning Associates survey and 1984/85 Malaysian Population and Fertility Survey.	MFLS-2 fetal mortality rates for pregnancies between 1970 and 1973 were higher than in MFLS-1, but lower than in FFPA survey (see text). Comparison with 1984/5 Malaysian Population and Fertility Survey shows New Sample especially under-reports induced abortions.	Use alternative methods of data collection for highly sensitive topics.	12
		MFLS-2 New	Expected relationship with other variables.	From 1950-85, period marked by declining infant, child, and maternal mortality, MFLS-2 implies rising fetal mortality rates (from 5% 1950 to 10% in 1985), opposite of expected trend. More educated women reported higher levels of miscarriage than less educated women. Conclude that miscarriages that happened in earlier periods and miscarriages to less educated women are underreported.		9
	Reporting approximate instead of exact birthweight	MFLS-1	Completeness of reports.	Approximate birthweight more likely to be reported by less educated women, ethnic Malays and urban residents; also more likely if birth occurred in other than clinic/hospital, child died in first year, involved longer recall period, or baby normal weight.	Pool approximate and exact reports to avoid serious sampling biases. Be sensitive to possible cultural and communication differences between interviewer and respondent in interviewer training.	1
Birthweight		MFLS-1	Comparison with vital statistics.	Distribution of exact birthweights and all birthweights (imputing values for approximate birthweights) for infants born between 1970-76 in MFLS-1 comparable in level and shape to 1977 vital statistics distribution.	Respondents should be asked to provide approximate values if they are unable to provide exact values.	1
		MFLS-1	Expected relationship with other variables.	Approximate "low" and "very low" birthweights associated with higher infant mortality.	Collect approximate reports if respondent unable to provide exact information.	3
		MFLS-2 New	Comparison with vital statistics and expected relationship with other variables.	Distribution of exact weights for births between 1982-86 in MFLS-2 shows approximately the same shape (normal distribution) as the 1984 vital statistics weights. "Low and very low " approximate weights highly predictive of infant mortality.	Collect approximate reports if respondent unable to provide exact information.	12

Topic		MFLS data set and sample	Method of evaluation	Findings	Implications	*
		Matched sample	Matched aggregate comparison and 1:1 match	Small but significant differences in aggregate mean weight (lbs). In 1:1 match, <0.5 lb. difference in reports for 2/3 births with exact weight. When differ, tendency is toward lower weight in MFLS-2 due to rounding down to nearest pound in MFLS-2. In 1976, 2/3's of birth weights had pounds and ounces; in 1988, only 50% had reports on non-zero ounces.	Collect birthweight information using meaningful categories: I.e. was birthweight <= 5.5 pounds.	10
	Birthweight<= 5.5 lbs.	Matched sample	Matched aggregate comparison and 1:1 match	No statistically significant differences in % reported <= 5.5 lbs in aggregate. In matched sample, 91% agreement.	For analytical purposes, construct meaningful categorical or discrete variables from amount reports.	10
Breastfeeding and post-partum amenorrhea	Whether child breastfed.	MFLS-1 (rounds 1 and 2)	1:1 match	Test-retest reliability ratio for whether ever breastfed first child is 0.91.		3
		MFLS-2 New	Comparison with 1984/5 Malaysian Population and Fertility Survey.	MFLS-2 contains a higher proportion of women who breastfed their youngest child (80%) compared with the MPFS (77%).	To improve reporting of retrospective events that are socially acceptable or unacceptable at time of interview, use probe or cross-reference with other events. E.g. if long period during which contraception not reported and no pregnancies occur, can probe about breastfeeding if not reported.	12
		MFLS-1 & MFLS-2 New; Matched sample	Matched aggregate comparison and 1:1 match	Breastfeeding rates by year of birth generally lower in MFLS-1 than in MFLS-2 New sample, though similar trend evident in both. Kappa coefficient for reports of ever breastfeeding for pre-1977 births reported by matched sample is 0.70 (excellent agreement). MFLS-1 qualified its initiation of breastfeeding question with: "even if it was for a few days" whereas MFLS-2 instead said "even if you just tried once or twice.	Maintain question wording precisely across waves to minimize possibility that changes over time reflects measurement error.	2

Topic		MFLS data set and sample	Method of evaluation	Findings	Implications	*
	Breastfeeding duration	MFLS-1 (rounds 1 and 2)	1:1 match	Correlation in breastfeeding duration reported in Round 1 and 2 was 0.91. Ethnic Malays, less educated respondents and longer recall periods positively related to reporting of "peak values."	Be more sensitive to cultural and communication differences between interviewer and respondent in interviewer training. Restrict asking about length of breastfeeding to most recent births or use calendar method to relate age at weaning to timing of other events to reduce reporting error.	4
		MFLS-1, MFLS-2 New, Matched sample	Matched aggregate comparison and 1:1 match	Heaping occurs at six month intervals. Differences in reports of durations between MFLS-1 and MFLS-2 were often large. Often, reported durations moved to or beyond the next heaping point.	Heaping can bias downward estimated effects of demographic variables on duration of breastfeeding.	4,6
		Matched sample	Matched aggregate comparison and 1:1 match	Significant difference in mean duration in aggregate. Nearly 2/3 reported durations differed by more than a month; nearly 50% by more than 3 months difference. More recent births show greatest differences.		10
		MFLS-1	Expected relationship with other variables.	Duration of post partum anovulation and duration of breastfeeding relationship are related as anticipated by physiological knowledge.		5
				Waiting time to conception in months, given the median duration of breastfeeding, is comparable to the association obtained in other settings.		15
	Breastfed < 4 months	Matched sample	Matched aggregate comparison and 1:1 match	No significant differences in aggregate level of breastfeeding < 4 months.		10
		Matched sample	1:1 match	Excellent agreement (K^2=0.68), though women significantly more likely to report "breastfed for at least four months" in MFLS-2 than in MFLS-1.		2
	Reasons for no or short breastfeeding	Matched sample	1:1 match	There is good agreement (K^2=0.41) of general reason for no or short breastfeeding. Reasons linked to specific events-child dying or mom returning to work-are reported more consistently than those where rationale is less salient or socially acceptable.	Use self-reported reasons for no or short breastfeeding rather than deriving reason based on self-reported behaviors.	7

Topic		MFLS data set and sample	Method of evaluation	Findings	Implications	*
			Comparison with other survey data.	Consistent with previous studies, no/insufficient breast milk was most commonly reported reason for not breastfeeding or for short duration of breastfeeding. Self-reported reasons do not appear to be less flawed than "objective" reports of breastfeeding initiation and duration.		7
	Type of 1st supplementary food was milk	MFLS-1 Rounds 2 and 3	1:1 match	Agreement excellent for first milk (Kappa=0.73). Education, Chinese or Indian ancestry, and recency of birth positively associated with reliability.	Be more sensitive to cultural and communication differences between interviewer and respondent in interviewer training.	4
	Post-partum amenorrhea (duration)	MFLS-1	Heaping.	Serious heaping on six-month intervals. Less educated and ethnic Malays more likely to heap; longer recall period increases heaping.	Examine sensitivity of results to which grouping heaped category is included in.	3
Education	Years completed education	MFLS-1	Comparison with census.	Women's level of education in 1957 and 1970 highly comparable to census data.		3
		MFLS-2 New	Expected relationship with other variables and comparison with census data.	Older persons reported less education, consistent with expectations. Percent of women who reported they had ever attended school by age group, in 1970 and 1980, comparable to census data, with some overstatement among older women.		12
	School attendance at age 15	MFLS-1	1:1 match	Retest reliability ratio for answer "Were you in school at age 15" compared with answer implied in Round 1 life history is low: 0.11. The life history portion of the survey was intended to include all events (including continuing schooling) from age 15 or since first marriage (whichever came first). The specific events to be included in the life history were not made explicit to respondent.	Make sure screener questions are not confusing or ambiguous.	3
Earnings		MFLS-1	Expected relationship with other variables and comparison with economic data.	From 1950-73, men's reported earnings histories yield income growth rate of 4.3% compared to GDP growth rate of 4.4%. Correlates of retrospective earnings similar to correlates reported in the literature for cross-sectional data.		13

Topic		MFLS data set and sample	Method of evaluation	Findings	Implications	*
		MFLS-2 New	Expected relationship with other variables.	Reporting error variance associated with retrospective reports of earnings rise with length of recall period.	Account for reporting error variance in multivariate models based on retrospectively reported data.	8
Housing characteristics	Electric lighting	MFLS-1	Comparison with census.	50% of MFLS households reported having electricity in the house where they lived in 1970 compared with 44% of households in 1970 census.		3
	Sanitation	MFLS-1	Comparison with census.	24% of MFLS households reported non-flush or flush toilet in the house where they lived in 1970 compared with 20% in 1970 census.		3
	Water supply	MFLS-1	Comparison with census.	48% of MFLS households reported piped water in the house where they lived in 1970, exactly the same as the percentage in the 1970 census.		3
Place of residence and migration history	Residence at birth and age 15	Matched sample	1:1 match	Over 90% of respondents place themselves at birth and age 15 with same or adjacent district in MFLS-1 and MFLS-2.		14
	Moves since age 15	Matched sample	1:1 match	Higher consistency for more educated respondents and for moves that are long-distance, occurred closer to time of interview, crossed district boundary, made jointly by spouse, noncircular or were coincident with other salient life events.	Use calendar method to link events to salient life events. Consider restricting analysis to most recent period when less salient moves most likely to accurately be recalled.	14
	Discrepancy in dates provided for moves	Matched sample	1:1 match	Moves are telescoped. Smaller discrepancy for younger, more educated respondents; for more recent moves, if salient coincident event, fewer total # moves (women); if interviewer judged responses as very reliable (men).	With panel design, use anchoring based on earlier reports to avoid telescoping of events. Use calendar method to link events to salient life events to aid respondents in determining the timing of events.	14

* References:

1. DaVanzo et al. 1984
2. DaVanzo et al. 1994
3. Haaga 1986
4. Haaga 1988
5. Habict et al. 1985
6. Klerman 1995
7. Kuate Defo and DaVanzo 1996
8. Lillard and Kilburn 1995
9. Panis and Lillard 1994
10. Peterson 1993
11. Sine 1993
12. Sine and Peterson 1993
13. Smith 1983
14. Smith and Thomas 1997
15. VanLandingham 1993