Characteristics Associated with Reliability in Reporting of Contraceptive Use: Assessing the Reliability of the Contraceptive Calendar in Seven Countries

Philip Anglewicz, Dana Sarnak, Alison Gemmill, and Stan Becker

Although the reproductive calendar is the primary tool for measuring contraceptive dynamics in low-income settings, the reliability of calendar data has seldom been evaluated, primarily due to the lack of longitudinal panel data. In this research, we evaluated the reproductive calendar using data from the Performance Monitoring for Action Project. We used population-based longitudinal data from nine settings in seven countries: Burkina Faso, Nigeria (Kano and Lagos States), Democratic Republic of Congo (Kinshasa and Kongo Central Provinces), Kenya, Uganda, Cote d'Ivoire, and India. To evaluate reliability, we compared the baseline cross-sectional report of contraceptive use (overall and by contraceptive method), nonuse, or pregnancy with the retrospective reproductive calendar entry for the corresponding month, measured at follow-up. We use multivariable regressions to identify characteristics associated with reliability or reporting. Overall, we find that the reliability of the calendar is in the "moderate/substantial" range for nearly all geographies and tests (Kappa statistics between 0.58 and 0.81). Measures of the complexity of the calendar (number of contraceptive use episodes, using the long-acting method at baseline) are associated with reliability. We also find that women who were using contraception without their partners/husband's knowledge (i.e., covertly) were less likely to report reliably in several countries.

Philip Anglewicz, Dana Sarnak, Alison Gemmill, Stan Becker, Department of Population, Family and Reproductive Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA. E-mail: panglew1@jhu.edu.

BACKGROUND

Although the reproductive calendar is the primary tool for measuring contraceptive dynamics in low- and middle-income settings, the validity and reliability of calendar data are largely unknown, primarily due to data limitations. A tool to collect a woman's recent pregnancy and contraceptive history, the reproductive calendar is a grid in which pregnancies, pregnancy outcomes (including births and pregnancy terminations), and contraceptive use (including method type and reason for discontinuation) are recorded for each calendar month over a two- to five-year period preceding the survey. Many large-scale surveys, like the Demographic and Health Surveys (DHS) and the National Survey of Family Growth (NSFG), have included reproductive calendars in their survey instruments for decades.

Data from the reproductive calendar are routinely used to calculate key family planning measures that inform policies worldwide. The reproductive calendar is used to calculate contraceptive discontinuation rates by method and reason (Ali, Cleland, and Shah 2012), as well as contraceptive switching and contraceptive failure rates (Polis et al. 2016), all of which are considered fundamental measures of a country's family planning program. Method discontinuation and switching, for example, are among the 18 core measures identified by FP2020 to evaluate countries' progress toward family planning goals and to measure the extent to which individuals' family planning needs are met (FP2020 2021). The importance of the reproductive calendar is widely acknowledged; Bradley et al. (2015) state that "Information collected in DHS calendars form the primary data source for the study of contraceptive use dynamics, particularly rates of contraceptive discontinuation, failure, and switching, in low- and middle-income countries" (pg. 21).

Despite the widespread reliance on the reproductive calendar to create these key family planning indicators, the validity and reliability of calendar data are largely unknown. Although the calendar has been implemented in more than 37 countries since 1990, yielding more than 200 total calendars (Bradley et al. 2015), few studies have examined the data quality of the calendar—and the ones that did have important limitations.

Among the few studies that have evaluated calendar data quality, most have used cross-sectional data (herein referred to as the "cross-sectional approach") (Bradley et al 2015; Becker and Diop-Sidibé 2003; Curtis and Blanc 1997; Becker and Sosa 1992; Westoff, Goldman, and Moreno, 1990; Goldman, Moreno, and Westoff 1989). For example, several studies evaluated calendar quality by calculating modern contraceptive prevalence rates (mCPRs) for each year available in the calendar and comparing calendar-based mCPRs to cross-sectional noncalendar-based mCPR estimates from DHS (Bradley et al. 2015; Strickler et al. 1997). Cross-sectional data have also been used to identify the extent of age heaping (Becker and Diop-Sidibe 2003), or for the population-level estimates described above, to compare the calendar with another mode for capturing contraceptive dynamics (Goldman, Moreno, Westoff 1989). Although these approaches provide useful information about calendar data quality at the population level, they do not permit the identification of individuals for whom the calendar is unreliable; and it is difficult to improve calendar data quality without knowing which women report unreliable calendar information. To measure the reliability of the calendar data, one would need longitudinal data for the same women, with a period of overlap in the calendar: as stated in Bradley et al. (2015) "An ideal way to assess the reliability

of retrospectively collected data would be to interview the same women multiple times" (pg. 21).

To date, there have been only three studies with the longitudinal design and calendar data necessary to achieve this objective (Callahan and Becker 2012; Strickler et al. 1997; Tumlinson and Curtis 2021), using data from only three countries: Kenya, Bangladesh, and Morocco. Of these studies, only one used nationally representative data; the Bangladesh study only included data for rural residents (Amin et al. 2010), and the Kenya study was among urban residents. All three studies involved populations with relatively high contraceptive use and low fertility compared to most countries in sub-Saharan Africa (SSA): the contraceptive prevalence rate was 44 percent in Morocco, 70 percent in Bangladesh, and 57 percent in Kenya (Callahan and Becker 2012; Strickler et al. 1997; Tumlinson and Curtis 2021). Only Tumlinson and Curtis (2021) used longitudinal calendar data from a country in SSA, many of which have among the highest fertility rates and lowest rates of contraceptive use in the world (United Nations 2020).

The three studies that have used longitudinal data to evaluate the reliability of calendar data generally agree on some features of the calendar: (1) the reliability of the calendar falls within the "moderate to substantial agreement" category (Kappa statistics between 0.41 and 0.80), and (2) women with more complex reproductive histories are less reliable in reporting their calendar information (Callahan and Becker 2012; Strickler et al. 1997; Tumlinson and Curtis 2021). However, there is also substantial disagreement and notable gaps in this research. First, there is variation in reliability across settings, with Kappa's ranging from 0.56 to 0.79 across countries and ways of testing reliability (Callahan and Becker 2012; Strickler et al. 1997; Tumlinson and Curtis 2021). Second, the characteristics associated with reliability are not consistent across studies: measures like age are not consistently associated with reliability across studies, and other measures, like household wealth and urban/rural residence, were not tested in all studies. Third, although all three studies hypothesize that the use of long-acting reversible methods (LARCs) is associated with greater reliability, this was only found in one (Tumlinson and Curtis 2021), potentially due to small sample sizes of LARC users in the other two studies.

The limitations of previous research are well-documented. Bradley et al. (2015) note that "Few studies to date have examined the quality of the contraceptive information collected in DHS calendars" (pg. 21), and that more has been done to examine quality at the population level instead of the level of the individual woman. Similarly, Callahan and Becker (2012) state that "The reliability of these methods for capturing accurate contraceptive histories over time…remains largely unknown" (pg. 3). Although the previous three studies were conducted in settings with varying sociodemographic and family planning characteristics, only one took place in SSA; the reliability of calendar data is otherwise unknown for these populations.

These three studies also included a limited range of factors that influence reliability. Previous studies have focused on demographic characteristics (e.g., age, education, and household income) and calendar complexity (such as the number of use episodes, parity, and method type). While these factors are indeed important influences on reliability, the underlying assumption in selecting these measures is that reliability is primarily influenced by the ability to remember the timing and order of reproductive events. But this misses another important potential influence on calendar reliability: women's willingness to

self-report reproductive events. Reporting contraceptive use, for example, may be sensitive in some contexts, and sensitive questions are often subjected to relatively larger biases in response and may, therefore, be less reliable (Bignami-VanAssche 2003; Knodel and Piampiti 1977). Therefore, there is a need to include measures that capture women's willingness to report contraceptive use in assessments of calendar reliability. We hypothesize that reliability is influenced both by the woman's ability to remember events (captured by demographic characteristics and calendar complexity), and her willingness to report various reproductive behaviors, which is influenced by factors like social norms and communication about contraceptive use with a husband or partner.

In this research, we used rarely available longitudinal panel data to evaluate the reliability of the reproductive calendar. In doing so, we addressed the prominent limitations of previous studies: we used these longitudinal panel data with the overlap in the reproductive calendar for the same women, which permitted the individual-level analysis that has been lacking. We expand the evidence base on how characteristics of women may affect their ability to retrospectively remember and report reproductive events, and examine whether calendar reliability is influenced by a woman's willingness to self-report reproductive behavior by testing whether sensitive contraceptive behaviors or norms around family planning are associated with reliability in reporting. We used recent data from seven countries that vary substantially in sociodemographic, fertility, and family planning profiles: Kenya, Burkina Faso, Nigeria, the Democratic Republic of Congo (DRC), Uganda, India, and Cote d'Ivoire.

METHODS

Data

For this study, we used data from the Performance Monitoring for Action Project (PMA). PMA started in 2013, with the goal of collecting representative data on key family planning indicators in Africa and Asia. To date, PMA has operated in 11 countries, collecting representative data at the national and/or subnational level. PMA uses a multistage stratified cluster design to draw a probability sample of households and females of childbearing age.

For data collection, PMA employs Resident Enumerators (REs), or women who live within or nearby the enumeration areas (EAs) where data are collected. Analysis suggests that this approach yields better data quality than interviewers who are not from the study area (Anglewicz et al. 2019; Safi 2019). Each RE is assigned to an EA of approximately 200 households. Data collection begins with mapping and listing of all households and health facilities in the EA, after which approximately 35 households are randomly selected. For selected households, the RE first administers a household survey that measures household assets, followed by a survey to all women aged 15–49 within the household that captures family planning-related behaviors. Data are collected on smartphones using Open Data Kit (ODK) as the program for data collection. After the interview is completed, the RE submits the data to a cloud server; these data are aggregated and downloaded by the PMA data management team for regular checks of data quality. Survey instruments are available on the PMA website (at https://www.pmadata.org/data/survey-methodology) and more information about the PMA

study design, data quality, data collection approach, and data use can be found in Zimmerman et al. (2017).

In its most recent iteration, starting in 2019, PMA shifted study designs from a repeated cross-sectional design to a longitudinal panel design. PMA collects representative longitudinal panel data in the seven countries listed above; PMA is nationally representative in Kenya, Burkina Faso, Niger, Uganda, and Cote d'Ivoire; and collects representative data at the subnational level in India (Rajasthan), Nigeria (Kano and Lagos states only), and DRC (Kinshasa and Kongo Central provinces only).[1] The countries included in our study vary in key family planning and fertility characteristics: the mCPR ranges from 8.1 percent to 44.2 percent, and long-acting contraceptive method prevalence ranges from 2.6 percent to 31.0 percent (PMA 2022). This variation is valuable because the reliability of the reproductive calendar could largely be a product of the contraceptive method mix and fertility profile. Previous research has shown that the reliability of reproductive calendar data is a function of the number of contraceptives used, the type of method used, and the number of pregnancies (Callahan and Becker 2012; Strickler et al. 1997; Tumlinson and Curtis 2021). Therefore, it is critical to have variation in fertility levels, contraceptive methods mix, and other family planning characteristics to adequately capture a range of aspects associated with the reliability of the reproductive calendar, and how these characteristics vary by setting.

For this analysis, we use data from the baseline (Phase 1) and follow-up (Phase 2) survey from nine settings in seven countries, which were collected between November 2019 and January 2021. The analytical sample in this study includes all women in both the Phase 1 (P1) and Phase 2 (P2) panels (e.g., women who were relocated and interviewed) for the nine geographies.[2] PMA has experienced exceptionally low attrition, obtaining over 70 percent of the baseline sample in all geographies (Appendix Table 3).

Measures and Analytic Methods

For this analysis, we measured reliability by comparing two different approaches to measuring contraceptive use, from (1) the main survey, which we call the "cross-sectional measure" and (2) the reproductive calendar, which provides a retrospective measure. For the former, PMA measures current contraceptive use by asking women if they or their partners are currently doing anything to delay or avoid getting pregnant, followed by a probe about coital-specific and traditional contraceptive method use, phrased as "Just to check, are you or your partner doing any of the following to avoid pregnancy: deliberately avoiding sex on certain days, using a condom, using withdrawal or using emergency contraception?" Following these questions, women are asked "Which method or methods are you using?," with all contraceptive methods listed as options. Women can list more than one current method; in the case of multiple

---

1 PMA also collects nationally representative data in Niger, but the second phase of these data was not available at the time of analysis, so we limited our study to the remaining seven countries.

2 Two ODK programming errors resulted in specific samples of women not being required to submit a contraceptive calendar at the end of the survey. First, REs interviewing panel women who reported never having used contraception and who did not have an intervening pregnancy between P1 and P2 were not required to submit a contraceptive calendar—this error was present for all countries. Second, REs interviewing panel women who were using family planning at the time of the P2 survey, who started within the calendar period, and who had no pregnancies during the calendar period were also not required to submit a contraceptive calendar—this error was present only for Kenya, Nigeria, DRC, and Burkina Faso. Therefore, these women were excluded from the present study. Exclusions range from 2 percent of the panel sample in Rajasthan to 13 percent of the panel sample in Nigeria Kano. We consider these biases in the interpretation of our results.

methods, PMA uses the most effective method. Women are also asked if they are currently pregnant. These questions are used for the P1 cross-sectional current reproductive status, in which women are categorized as nonusers, pregnant, or users of a specific method.

PMA's method for completing the reproductive calendar is like that of the DHS and NSFG. Due to the complexity of the calendar information, REs used a paper aid for data collection. The paper aid includes up to 36 boxes (each box representing one month of time, so up to three years) divided into three sections (each representing 12 months of time) in which to record information about the woman's experiences with pregnancy and contraceptive use. The calendar includes two columns, the first captures pregnancies, live births, pregnancy terminations (miscarriages or abortions), and contraceptive use (by method type), and the second captures contraceptive discontinuation by reason. The RE starts by recording information about pregnancies, terminations, and births, which serve as "anchors" for the subsequent recording of contraceptive use and discontinuation. After this information has been added to the calendar paper aid, the RE reviews for coherence and probes for more information if necessary, and then transfers the information from the paper aid to the ODK phone survey. After completing the interview, the REs took pictures of the paper and pencil calendar, so that the data management teams could compare the picture with the data entered in the phone. In doing so, we found very minimal errors between the paper form and the data.

Between these two approaches, we believe that the cross-sectional measure is likely more accurate than the retrospective measure from the contraceptive calendar. Unlike the retrospective calendar approach, the cross-sectional measure does not require remembering contraceptive behaviors from the past and is likely to be more accurate as a result (Tsui et al. 2021).

To evaluate the reliability of women in their reporting of their reproductive status, we compared the two measures above: the report of contraceptive use or pregnancy from the P1 survey with the status reported for the same month from the woman's P2 retrospective reproductive calendar. We used two different approaches to compare these measures. First, we examined the concordance of women in reporting a three-category measure: nonuse, pregnancy/pregnancy outcomes, and contraceptive use. The total percent agreement represents the diagonals of the 3×3 cross tabulation of these categories divided by the total sample size. This category of concordance is herein referred to as "3×3 concordance." Prior calendar evaluations (Strickler et al. 1997; Tumlinson and Curtis 2021) have used this approach to compare the reliability of cross-sectional reporting with retrospective calendar reporting. Second, we examined the concordance of women in reporting a 10-category measure, which includes nonuse, pregnancy/pregnancy outcomes, and specific contraceptive method use (male/female sterilization, intrauterine device (IUD), implant, injectable, pill, condom, other modern, and traditional). The total percent agreement represents the diagonals of the 10×10 cross tabulation of these categories divided by the total sample. This category of concordance is herein referred to as "10×10 concordance." This category of concordance adds a level of nuance to the 3×3 method; in this measure, the level of agreement also requires women to be consistent in the reporting of their specific method, which is not captured in the 3×3.

Next, we examined the percent agreement and Kappa statistics for the two measures of concordance outlined above, the 3×3 and the 10×10. We focus on the row totals, that is,

the percent of women who reported each status in their P2 retrospective calendar for the P1 survey month out of the total number of women who reported that status in the P1 survey. The total percent agreement was calculated as the total number of women who were concordant across the categories out of the total number of women. Kappa statistics were also computed as a means to evaluate how likely the concordance departs from chance (Landis and Koch 1977).

In addition to the analysis of agreement in the exact month, we also examined agreement within a +/− two-month period. The results, in Appendix Table 2, show that the percentage of agreement increases when the time period is expanded. But the change in agreement varies across countries, with the smallest overall increase in agreement was in Rajasthan (where the reliability was already high) and greatest in Kano (a setting with relatively lower reliability in terms of the Kappa statistic). After expanding the reporting period by +/− two months, there is still a percentage that does not agree, which suggests that there are events that are misreported as opposed to just inaccuracies in timing.

Third, we conducted multivariable logistic regression to determine which measures were associated with reliability. The outcome variable, whether a woman was concordant in her reporting, comes from the 10×10 table, in which she was considered concordant if she consistently reported one of the 10 following statuses for the P1 survey month in both surveys: nonuse; pregnancy/birth outcome; male/female sterilization, IUD, implant, injectable, pill, condom, other modern, and traditional.

In our multivariable regression analysis, we included P1 sociodemographic covariates: age of the woman (15–24 years; 25–34 years; 35 years or older); parity (0 births, 1–2 births, 3–4 births, 5+ births); wealth quintile; the highest level of schooling (none/primary; secondary or higher); and residence (urban/rural). Like previous studies, we examined how reliability varied by the complexity of the P2 calendar using three measures: the number of pregnancies reported in the P2 calendar (none, 1 pregnancy, 2+ pregnancies); the number of contraceptive use episodes reported in the P2 calendar (none, 1 episode, 2+ episodes); and a binary variable for if a long-acting method was reported at P1 (no method reported/short-acting method at P1; long-acting method reported at P1).

We also explored the factors that might influence women's willingness to report contraceptive use, focusing on three: covert contraceptive use, and two measures of contraceptive use norms. For covert use, in each survey wave, women who reported currently using a female-controlled method (i.e., any method except for male condom, withdrawal, or male sterilization) were asked if their partner knew about the use. Women who were using a female-controlled method and who responded that their partners did not know about their use were classified as covert users; all other users were classified as overt users. To examine whether covert use of contraception was associated with the reliability of reporting, we added a binary measure of covert use at either P1 and/or P2 to the multivariable regression models described above. Women were only included in this analysis if they had a value for this measure in either wave; therefore, this analysis is limited to women who were using a female-controlled method in either P1 or P2. The two measures of contraceptive use norms are (1) the P1 EA-level average prevalence of contraceptive use, and (2) the P1 EA-level average level of agreement with the statement that "Family planning is for married women only." We expect that higher levels of contraceptive use in the community represent a favorable

environment for contraceptive use and will be associated with greater reliability; while a higher percentage of community agreement with the statement that family planning is only for married women represents a more restrictive environment and will be associated with less reliability. We conduct separate multivariable regressions for each of these three measures and include all demographic and calendar complexity measures as well. Because the two norm measures are captured at the community (EA) level, we use multilevel regression models.

PMA weights all data to account for study design and nonresponse, and attrition for the longitudinal panel data. In all analyses, we used design-based logistic regressions to adjust for design effects (DEFF>1) due to the multistage stratified cluster design. PMA has consistently obtained a response rate of greater than 70 percent in all settings. We also compared characteristics between women who were interviewed in both P1 and P2, and those lost to follow-up. As expected, the lost to follow-up were younger, better educated, and have fewer children (shown in Appendix Table 3). To address the differential loss to follow-up, we created inverse probability weights by first identifying sociodemographic characteristics associated with the likelihood of reinterview in P2 (at p<0.05), then calculated the predicted values of these measures, and finally created weights that were the inverse of the predicted values.[3] For our regression results, we show odds ratios and 95 percent confidence intervals; results that are statistically significant at p<0.05 are in bold font.

RESULTS

This study benefits from data across a range of settings that vary in sociodemographic characteristics and contraceptive use patterns. As shown in Table 1, while age distributions were generally similar across sites, there was more variation in parity, with at least one-third of women nulliparous in Nigeria-Lagos and DRC Kinshasa, while in Nigeria-Kano, 41 percent of women had given birth to 5+ children. Similarly, the highest schooling level attained varied across sites; most women in Nigeria-Lagos and DRC Kinshasa had at least a secondary education (88 percent and 92 percent, respectively). On the other hand, close to two-thirds of women in Nigeria-Kano, Burkina Faso, and Cote d'Ivoire had no formal education or only primary education (64 percent, 76 percent, and 64 percent, respectively). Most of the population lived in rural residences in Kenya, Nigeria-Kano, Burkina Faso, Uganda, and Rajasthan, but most of the population had urban residences in Cote d'Ivoire, and the samples were entirely urban in Nigeria-Lagos and DRC-Kinshasa.

Regarding contraceptive use patterns, P1 reproductive status differed considerably across sites in terms of contraceptive prevalence and method mix. Contraceptive use was lowest

---

3   We also conducted two additional analyses. First, to examine whether the total percent agreement was a function of the distribution of nonuse and FP use, we examined the total percent agreement between P1 current use and P2 calendar reporting among two subsamples of users. We compared the P1 current use status to the P2 calendar status of women who reported using male/female sterilization, implant, IUD, injectables, pills, condoms, other modern methods, or traditional methods at P1, which we herein name the "8×10 table." Second, to further explore the concordance of short-acting method user specifically, we also examined the percent agreement in reporting of short-acting users only. We compared the P1 current use status to the P2 calendar status of women who reported using injectables, pills, condoms, other modern methods, or traditional methods at P1. We call this type of agreement "5×10 agreement." Results of these analyses (not shown) did not change the substantive interpretation of our findings.

TABLE 1 Percent distribution of sociodemographic and calendar complexity characteristics for women in nine settings, Performance Monitoring for Action (PMA) Phase 2

Study site									
Sociodemographic group and calendar complexity	Kenya	Nigeria Lagos	Nigeria Kano	DRC Kinshasa	DRC Kongo Central	Burkina Faso	Uganda	Cote d'Ivoire	Rajasthan
Age									
15–24	35.2	24.5	40.5	38.5	33.5	35.0	40.5	36.4	32.3
25–34	32.6	31.2	30.4	31.0	32.0	30.5	31.7	31.8	32.3
35+	32.3	44.3	29.2	30.5	34.5	34.5	27.8	31.7	35.3
Number of children									
0 children	24.8	32.7	26.9	38.1	19.1	20.8	24.8	27.1	28.2
1–2 children	30.5	28.7	15.2	30.0	30.4	23.6	27.0	30.7	39.0
3–4 children	24.3	29.5	17.3	19.6	24.9	23.3	20.1	20.5	27.0
5+ children	20.3	9.2	40.5	12.3	25.6	32.3	28.1	21.7	5.8
Wealth									
Lowest quintile	19.8	17.0	17.4	14.6	18.2	21.1	19.4	17.0	15.5
Middle lowest quintile	23.1	19.2	19.7	19.9	19.1	19.9	18.2	16.9	16.9
Middle quintile	20.1	18.5	21.7	18.1	19.1	19.3	18.7	17.6	20.4
Middle highest quintile	17.4	22.0	19.3	22.7	21.9	19.2	19.0	22.5	22.9
Highest quintile	19.6	23.3	21.9	24.7	21.8	20.4	24.7	26.0	24.3
Education									
None/Primary	47.2	11.6	63.7	7.6	37.5	76.1	59.4	63.8	48.4
Secondary+	52.8	88.4	36.3	92.4	62.5	23.9	40.6	36.2	51.6
Urban/rural residence									
Rural	70.9	N/A	61.0	N/A	N/A	76.3	72.0	38.2	75.0
Urban	29.1	N/A	39.0	N/A	N/A	23.7	28.0	61.8	25.0
Contraceptive use status									
Nonuse	43.7	54.2	80.8	54.3	49.4	58.1	50.1	54.1	42.5
Pregnancy/birth	5.3	4.5	9.3	5.8	8.6	9.4	9.0	8.4	3.0
Female/male sterilization	2.3	0.3	0.2	0.3	0.9	0.1	1.7	0.0	33.1
Implant	17.1	4.9	2.7	6.8	10.5	13.5	9.4	6.4	0.0
IUD	1.4	2.1	0.0	0.2	0.1	1.5	1.7	0.7	1.1
Injectable	17.9	3.2	3.1	3.4	5.0	8.5	13.4	6.9	1.5
Pill	4.0	3.4	1.3	0.7	2.5	2.1	2.0	6.2	2.9
Condom	3.6	8.6	0.2	6.5	6.1	3.9	3.5	6.5	8.9
Other modern	1.6	3.0	0.0	5.7	2.0	0.3	2.4	1.9	0.9
Traditional	3.1	15.8	2.2	16.3	14.6	2.6	6.7	9.1	6.0
Pregnancies in P2 calendar									
0 pregnancies	68.6	73.4	50.9	70.1	57.1	55.4	56.5	64.3	80.5
1 pregnancy	27.7	21.9	46.3	25.9	37.2	39.9	36.7	32.0	16.8
2+ pregnancies	3.7	4.7	2.8	4.0	5.7	4.7	6.8	3.7	2.7

TABLE 1 (Continued)

Study site

Sociodemographic group and calendar complexity	Kenya	Nigeria Lagos	Nigeria Kano	DRC Kinshasa	DRC Kongo Central	Burkina Faso	Uganda	Cote d'Ivoire	Rajasthan
Use episodes in P2 calendar									
0 use episodes	36.8	49.3	87.8	42.0	47.0	53.6	40.9	52.8	40.9
1 use episode	42.0	32.9	10.1	36.8	37.2	34.9	40.2	35.6	52.3
2+ use episodes	21.2	17.9	2.1	21.2	15.8	11.5	18.9	11.6	6.8
N=	6,587	1,037	864	1,812	1,370	4,985	2,962	2,909	4,435

NOTE: All estimates are weighted for study design characteristics and loss to follow-up using inverse propensity weighting techniques. Nigeria-Lagos and DRC-Kinshasa sites are urban sites and the sampling frame for DRC-Kongo Central did not include an urban/rural distinction, therefore, these are missing the urban/rural residence categories; not all categories add up to 100%, due to rounding.

in Nigeria-Kano, where only 9 percent of the sample reporting current use; use was highest in Kenya (51 percent) and Rajasthan (45 percent). Injectables and implants were the most popular methods in Kenya, Burkina Faso, and Uganda. Traditional methods were the most reported in Nigeria-Lagos, DRC, and Cote d'Ivoire. Female sterilization was the most commonly reported method in Rajasthan. At least 50 percent of women reported no pregnancies in their P2 retrospective calendar; yet, the number reporting at least one pregnancy in their P2 calendar ranged from 20 percent in Rajasthan to 49 percent in Nigeria-Kano. The number of use episodes was highest in Kenya, where 63 percent reported at least one use episode, and lowest in Nigeria-Kano, where 89 percent of women reported no use episodes.

Table 2 shows the results of the 3×3 concordance analysis, which is the total agreement with the three-category measure of nonuse, pregnancy, and contraceptive use. We show the percentage of agreement for each category, along with the total percentage of agreement and the Kappa statistic. Overall, we find that agreement was similar across countries, ranging from 78 percent in Uganda, Cote d'Ivoire, and DRC Kongo to 90 percent in Rajasthan. Kappa statistics showed slightly more variation, ranging from 0.60 in Cote d'Ivoire to 0.81 in Rajasthan; though all fell in the "moderate/substantial" agreement category (i.e., Kappa statistics between 0.41 and 0.80), except for Rajasthan, which would be considered to have "excellent reliability" (Landis and Koch 1977). When we examine this by specific status (nonuse, pregnancy, and contraceptive use), we find more variation: across most sites, the percent agreement was generally higher for nonuse (81.3–90.4 percent) and pregnancy (75.7–95.2 percent) but lowest for FP use (60.9–91.6 percent), except for Rajasthan, where all three statuses had 90–92 percent agreement.

Once we disaggregate the contraceptive use into specific method types (Table 3), we generally find lower percentages of agreement and lower Kappa statistics. The total percent agreement decreased in all countries from those shown in Table 2, ranging from 73 percent in DRC Kongo to 86 percent in Rajasthan. This pattern was similar for Kappa statistics, which ranged from 0.57 in Cote d'Ivoire to 0.80 in Rajasthan. Across sites, the percent agreement was typically the highest for long-acting methods (female/male sterilization, implant, and IUD), less so for short-acting methods, such as injectable and pill, and lowest for coital-specific methods like the male condom. The percent agreement of traditional method use was typically between pill/condom and injectable in reliability.

We show results from two additional and related analyses in the online Appendix (Table 1). When we examined the total percent agreement only among women who reported currently using at P1, the percent agreement of their specific method was lower at all sites than the full sample of women (either 3×3 or 10×10), and more across-country variation is evident: just over half of women in Cote d'Ivoire and Nigeria-Kano were consistent in their method reporting, compared to 61 percent of women in Nigeria-Lagos and DRC-Kongo Central. Rajasthan had the highest percentage, at 84 percent of agreement among users. We further examined the reliability of reporting among women using short-acting methods (injectable, pill, condom, other modern, and traditional) at P1. Among women who reported P1 current use of any of these short-acting methods, total agreement decreased across all sites compared to other tests of agreement. Around 50 percent of women in Nigeria-Kano, DRC Kongo Central, Uganda, and Cote d'Ivoire reliably reported their short-acting methods, and the highest was 75 percent of women in DRC Kinshasa who reported these methods reliably.

TABLE 2   Percent agreement and kappa statistics for concordance in reporting of reproductive status in the same reference month (the 3×3 Concordance Analysis), 2019–2021 Performance Monitoring for Action data from nine geographies

Kenya	% Agreement/Kappa	Unweighted n
Nonuse	81.3	2,961
Pregnancy outcome	93.1	360
FP use	83.7	3,266
Total % agreement	82.5	
Kappa statistic	0.69	
Unweighted n=		6,587

Nigeria-Lagos	% Agreement/Kappa	Unweighted n
Nonuse	86.3	553
Pregnancy outcome	86.7	56
FP use	73.9	428
Total % agreement	81.6	
Kappa statistic	0.66	
Unweighted n=		1,037

Nigeria-Kano	% Agreement/Kappa	Unweighted n
Nonuse	90.4	637
Pregnancy outcome	75.7	110
FP use	60.9	117
Total % agreement	85.2	
Kappa statistic	0.65	
Unweighted n=		864

DRC-Kinshasa	% Agreement/Kappa	Unweighted n
Nonuse	87.0	876
Pregnancy outcome	95.2	113
FP use	84.5	823
Total % agreement	86.7	
Kappa statistic	0.77	
Unweighted n=		1,812

DRC-Kongo Central	% Agreement/Kappa	Unweighted n
Nonuse	83.4	632
Pregnancy outcome	80.7	133
FP use	71.8	606
Total % agreement	78.2	
Kappa statistic	0.63	
Unweighted n=		1,371

Burkina Faso	% Agreement/Kappa	Unweighted n
Nonuse	85.5	2,626
Pregnancy outcome	94.4	413
FP use	77.0	1,859
Total % agreement	82.7	
Kappa statistic	0.70	
Unweighted n=		4,898

Uganda	% Agreement/Kappa	Unweighted n
Nonuse	82.2	1,611
Pregnancy outcome	87.2	288
FP use	70.2	1,063
Total % agreement	78.4	
Kappa statistic	0.62	
Unweighted n=		2,962

Cote d'Ivoire	% Agreement/Kappa	Unweighted n
Nonuse	82.4	1,763
Pregnancy outcome	81.5	284
FP use	69.8	862
Total % agreement	78.3	
Kappa statistic	0.60	
Unweighted n=		2,907

(Continued on next page)

TABLE 2   (Continued)

Kenya	% Agreement/Kappa	Unweighted n
Rajasthan	**% Agreement/Kappa**	**Unweighted n**
Nonuse	89.7	1,879
Pregnancy outcome	91.5	201
FP use	91.6	2,355
Total % agreement	89.5	
Kappa statistic	0.81	
Unweighted n=		4,435

NOTE: Agreement is defined as the weighted percent of women who reported the respective reproductive status at the P1 survey and also reported the same status in P2 calendar for the same month of P1 survey.

The results of our multivariable analysis are shown in Table 4, which includes sociodemographic and calendar complexity measures that are associated with reliability. Several results are consistent across geographies: across all sites except Nigeria-Kano, women who reported using a long-acting method at P1 had greater odds of being reliable in their report than women not using long-acting methods in P1. In all sites except Nigeria-Kano and Rajasthan, women with more use episodes had lower odds of concordant reporting. Across most sites, higher parity had a negative relationship with concordance (except Nigeria-Lagos, DRC-Kinshasa, and Uganda).

Other results were consistent in a subset of geographies. In DRC-Kongo, Cote d'Ivoire, and Rajasthan, older age was associated with higher odds of being concordant. In another three sites, Nigeria-Kano, DRC-Kongo, and Burkina Faso, higher education was associated with lower concordance. Also, having more pregnancies in the P2 calendar was associated with lower odds of concordance in Kenya, Nigeria-Kano, and Burkina Faso. Finally, we see that the relationship with wealth varies across settings: increased wealth was associated with higher concordance in Nigeria-Kano and DRC Kongo but lower concordance in Cote d'Ivoire.

Finally, in Table 5, we see the results for the association between calendar reliability and covert use, and the two measures of contraceptive use norms: (1) the EA-level average prevalence of contraceptive use, and (2) the EA-level average level of agreement with the statement that "Family planning is for married women only." We find that reporting of covert use at either P1 or P2 was associated with lower odds of reporting use reliably in six of the eight settings where we tested this relationship (at p<0.10): Kenya, Nigeria-Lagos, DRC-Kinshasa, DRC-Kongo Central, Cote d'Ivoire, and India-Rajasthan (this was not tested in Nigeria-Kano due to the small number of female-controlled method users). The EA-level average prevalence of contraceptive use was associated with reliability in only two settings, Nigeria-Lagos and Nigeria-Kano, but the association was in opposite directions: in Lagos, there was a positive relationship between the prevalence of contraceptive use and reliability, while the relationship was negative in Kano. Finally, we found only one statistically significant association between reliability and the EA-level average agreement that family planning is for married women only, where a higher percentage of agreement was associated with lower odds of reliable reporting.

TABLE 3 Percent agreement and kappa statistics for concordance in reporting of reproductive status in the same reference month (the 10×10 concordance analysis), by contraceptive method use, 2019–2021 Performance Monitoring for Action data from nine geographies

	Kenya		Nigeria Lagos		Nigeria Kano		DRC Kinshasa		DRC-Kongo Central	
	% Agree-ment/Kappa	Unweighted n	% Agree-ment/Kappa	Unweighted n	% Agree-ment/Kappa	Unweighted n	% Agree-ment/Kappa	Unweighted n	% Agree-ment/Kappa	Unweighted n
Nonuse	81.3	2,961	86.3	553	90.4	637	87.0	876	83.4	632
Pregnancy outcome	93.1	360	86.7	56	75.7	110	95.2	113	80.7	133
F/M sterilization	82.8	170	100.0	3	100.0	2	85.2	9	85.7	8
Implant	85.4	1,179	84.0	55	61.3	35	94.5	132	88.3	134
IUD	70.2	115	90.8	20	40.8	2	100.0	4	100.0	4
Injectable	68.7	1,148	60.3	35	46.1	42	67.0	58	72.4	69
Pill	56.1	215	52.7	39	76.5	15	57.7	24	38.9	37
Condom	45.7	196	54.3	99	67.8	2	73.4	131	37.1	90
Other modern	32.0	88	29.1	39	0.0	2	72.7	93	55.8	52
Traditional	45.2	155	61.3	138	100.0	17	76.3	372	53.7	211
Total % agreement	75.8		75.9		84.5		83.4		72.6	
Kappa statistic	0.67		0.64		0.65		0.77		0.62	
Unweighted n=		6,587		1,037		864		1,812		1,370

	Burkina Faso		Uganda		Cote d'Ivoire		Rajasthan	
	% Agree-ment/Kappa	Unweighted n	% Agree-ment/Kappa	Unweighted n	% Agree-ment/Kappa	Unweighted n	% Agree-ment/Kappa	Unweighted n
Nonuse	85.5	2,626	82.2	1,611	82.4	1,763	89.7	1,879
Pregnancy outcome	94.4	413	87.2	288	81.5	284	91.5	201
F/M sterilization	56.8	6	90.3	46	0.0	1	97.5	1,386
Implant	81.2	740	78.2	293	80.6	161	100.0	1
IUD	92.1	89	79.6	48	94.0	9	73.6	75
Injectable	58.0	377	61.1	342	61.6	174	59.2	42
Pill	47.1	214	57.9	47	54.2	176	66.7	181
Condom	57.5	287	21.5	86	43.9	142	68.5	398
Other modern	41.5	15	30.1	53	13.4	43	53.1	45
Traditional	48.8	128	40.6	148	54.3	154	69.0	227
Total % agreement	79.1		74.6		74.3		86.0	
Kappa statistic	0.69		0.62		0.58		0.80	
Unweighted n=		4,895		2,962		2,907		4,435

NOTE: Agreement is defined as the weighted percent of women who reported the respective reproductive status at the P1 survey and also reported the same status in P2 calendar for the same month of P1 survey.

TABLE 4 Adjusted odds ratios and confidence intervals for factors associated with reliable reporting of reproductive calendar, 2019–2021 Performance Monitoring for Action data from nine geographies

Sociodemographic group and calendar complexity	Kenya aOR	Kenya 95% CIs	Nigeria Lagos aOR	Nigeria Lagos 95% CIs	Nigeria Kano aOR	Nigeria Kano 95% CIs	DRC Kinshasa aOR	DRC Kinshasa 95% CIs	DRC Kongo aOR	DRC Kongo 95% CIs
Age										
15–24 (ref)	1.00		1.00		1.00		1.00		1.00	
25–34	0.98	0.77 1.25	1.12	0.51 2.43	0.58	0.30 1.14	0.88	0.60 1.29	1.31	0.80 2.16
35+	0.82	0.63 1.06	0.91	0.40 2.08	0.70	0.33 1.46	0.83	0.49 1.39	**2.12**	**1.04 4.33**
Number of children										
0 children (ref)	1.00		1.00		**1.00**		1.00		1.00	
1–2 children	**0.46**	**0.32 0.64**	0.70	0.34 1.44	**0.03**	**0.00 0.21**	1.02	0.60 1.72	0.56	0.30 1.08
3–4 children	0.53	0.37 0.77	0.65	0.29 1.46	**0.02**	**0.00 0.21**	1.45	0.74 2.83	**0.28**	**0.12 0.64**
5+ children	0.54	0.36 0.81	0.45	0.19 1.06	**0.02**	**0.00 0.22**	1.40	0.66 2.95	**0.24**	**0.09 0.61**
Wealth										
Lowest quintile (ref)	1.00		1.00		1.00		1.00		1.00	
Middle lowest quintile	1.01	0.84 1.23	0.88	0.43 1.80	1.87	0.79 4.43	0.69	0.39 1.22	**1.85**	**1.15 2.97**
Middle quintile	1.16	0.94 1.44	0.85	0.43 1.67	1.31	0.86 2.00	0.94	0.53 1.65	1.81	0.95 3.46
Middle highest quintile	0.93	0.71 1.22	1.21	0.62 2.36	1.24	0.58 2.66	0.94	0.55 1.60	**1.94**	**1.20 3.13**
Highest quintile	1.11	0.83 1.49	0.75	0.35 1.65	**3.60**	**1.60 8.08**	1.29	0.65 2.55	**1.64**	**1.02 2.63**
Highest level of schooling										
None/Primary (ref)	1.00		1.00		1.00		1.00		1.00	
Secondary+	0.88	0.75 1.04	0.78	0.46 1.30	**0.24**	**0.12 0.48**	1.43	0.85 2.38	**0.57**	**0.41 0.78**
Urban/rural residence										
Rural (ref)	1.00				1.00					
Urban	0.93	0.75 1.15			0.90	0.66 1.22				
# Pregnancies in P2 calendar										
0 pregnancies (ref)	1.00		1.00		1.00		1.00		1.00	
1 pregnancy	**0.73**	**0.61 0.88**	0.75	0.46 1.21	**0.40**	**0.19 0.86**	0.81	0.50 1.31	0.93	0.59 1.48
2+ pregnancies	**0.60**	**0.40 0.91**	0.84	0.32 2.22	0.43	0.09 1.92	1.00	0.38 2.61	1.43	0.57 3.57
# Use episodes in P2 calendar										
0 use episode (ref)	1.00		1.00		1.00		1.00		1.00	
1 use episode	**0.30**	**0.23 0.40**	**0.29**	**0.17 0.47**	0.97	0.53 1.78	**0.26**	**0.16 0.42**	**0.27**	**0.15 0.46**
2+ use episodes	**0.23**	**0.17 0.30**	**0.18**	**0.11 0.31**	1.33	0.33 5.31	**0.12**	**0.07 0.21**	**0.27**	**0.14 0.52**
Long-acting method use at P1										
No use/SAM method (ref)	1.00		1.00		1.00		**1.00**		**1.00**	
Long-acting method	**3.06**	**2.34 3.99**	**4.23**	**1.76 10.18**	0.30	0.08 1.17	**6.31**	**2.40 16.62**	**5.20**	**2.47 10.96**

	Burkina Faso aOR	Burkina Faso 95% CIs	Uganda aOR	Uganda 95% CIs	Cote d'Ivoire aOR	Cote d'Ivoire 95% CIs	Rajasthan aOR	Rajasthan 95% CIs
Age								
15–24 (ref)	1.00		1.00		1.00		1.00	
25–34	1.10	0.74 1.64	0.97	0.47 1.99	1.10	0.84 1.44	1.08	0.68 1.71

TABLE 4   (Continued)

	Burkina Faso			Uganda			Cote d'Ivoire			Rajasthan		
	aOR	95% CIs		aOR	95% CIs		aOR	95% CIs		aOR	95% CIs	
35+	1.18	0.72	1.96	1.09	0.38	3.16	1.67	1.18	2.38	1.91	1.04	3.51
Parity												
0 children (ref)	1.00			1.00			1.00			1.00		
1–2 children	**0.38**	**0.25**	**0.57**	0.50	0.15	1.64	0.72	0.49	1.05	**0.05**	**0.02**	**0.13**
3–4 children	**0.34**	**0.21**	**0.57**	0.44	0.17	1.10	0.78	0.48	1.27	**0.04**	**0.01**	**0.13**
5+ children	**0.32**	**0.17**	**0.62**	0.44	0.12	1.61	**0.54**	**0.31**	**0.93**	**0.04**	**0.01**	**0.20**
Wealth												
Lowest quintile (ref)	1.00			1.00			1.00			1.00		
Middle lowest quintile	0.81	0.56	1.16	1.09	0.73	1.63	0.73	0.50	1.07	1.35	0.77	2.36
Middle quintile	0.89	0.58	1.36	1.06	0.73	1.54	0.80	0.55	1.15	1.75	0.86	3.56
Middle highest quintile	0.94	0.65	1.35	1.11	0.71	1.73	0.77	0.54	1.09	1.71	0.59	4.97
Highest quintile	0.77	0.52	1.14	1.06	0.31	3.63	**0.64**	**0.42**	**0.96**	1.84	0.81	4.17
Highest level of schooling												
None/Primary (ref)	1.00			1.00			1.00			1.00		
Secondary+	**0.71**	**0.52**	**0.98**	0.86	0.47	1.58	0.98	0.67	1.43	0.97	0.68	1.38
Urban/rural residence												
Rural (ref)	1.00			1.00			1.00			1.00		
Urban	0.88	0.66	1.17	0.72	0.38	1.36	0.95	0.68	1.31	0.70	0.34	1.46
# Pregnancies in P2 calendar												
0 pregnancies (ref)	1.00			1.00			1.00			1.00		
1 pregnancy	0.80	0.61	1.06	0.78	0.44	1.39	0.88	0.70	1.10	0.94	0.58	1.53
2+ pregnancies	**0.52**	**0.31**	**0.85**	0.63	0.20	1.99	0.81	0.31	2.13	1.17	0.60	2.26
# Use episodes in P2 calendar												
0 use episodes (ref)	1.00			1.00			1.00			1.00		
1 use episode	**0.43**	**0.31**	**0.60**	**0.45**	**0.21**	**0.94**	**0.29**	**0.24**	**0.35**	0.69	0.35	1.36
2+ use episodes	**0.49**	**0.30**	**0.82**	**0.37**	**0.14**	**1.02**	**0.27**	**0.17**	**0.44**	1.06	0.47	2.41
Long-acting method use at P1												
No use/SAM method (ref)	1.00			1.00			1.00			1.00		
Long-acting method	**1.92**	**1.28**	**2.86**	**1.89**	**1.02**	**3.49**	**2.62**	**1.18**	**5.82**	**10.73**	**3.76**	**30.63**

NOTE: Nigeria-Lagos and DRC-Kinshasa sites are urban sites; and the sampling frame for DRC-Kongo Central did not include an urban/rural distinction, therefore, these are missing the urban/rural residence categories. Short-acting methods include injectable, pill, condoms, other modern, and traditional methods. Long-acting methods include male/female sterilization, IUD, and implants. Boldfaced odds ratios and CIs are significant at p<0.05.

Abbreviations: aOR, adjusted odds ratios; 95% CIs, 95% confidence intervals.

TABLE 5   Adjusted odds ratios and confidence intervals for the effects of covert use and EA-level community norms on reliable reporting in the reproductive calendar for nine geographic regions, 2019–2021 Performance Monitoring for Action data

Setting and covariates	Adjusted Odds Ratios	95% Confidence Intervals	
Kenya			
Reported covert use at P1 or P2	**0.58**	**0.45**	**0.75**
EA-level CPR	0.96	0.48	1.92
EA-level % agree that family planning is for married women only	**0.72**	**0.53**	**0.98**
Nigeria-Lagos			
Reported covert use at P1 or P2	**0.51**	**0.25**	**1.04**
EA-level CPR	**3.03**	**1.05**	**8.72**
EA-level % agree that family planning is for married women only	0.73	0.37	1.47
Nigeria-Kano			
Reported covert use at P1 or P2	N/A	N/A	N/A
EA-level CPR	**0.12**	**0.01**	**1.07**
EA-level % agree that family planning is for married women only	1.74	0.61	4.95
DRC-Kinshasa			
Reported covert use at P1 or P2	**0.38**	**0.16**	**0.94**
EA-level CPR	0.24	0.03	1.80
EA-level % agree that family planning is for married women only	1.11	0.38	3.24
DRC-Kongo Central			
Reported covert use at P1 or P2	**0.29**	**0.11**	**0.82**
EA-level CPR	0.46	0.12	1.79
EA-level % agree that family planning is for married women only	1.04	0.52	2.08
Burkina Faso			
Reported covert use at P1 or P2	0.64	0.35	1.16
EA-level CPR	1.68	0.66	4.31
EA-level % agree that family planning is for married women only	0.83	0.42	1.63
Uganda			
Reported covert use at P1 or P2	0.39	0.13	1.22
EA-level CPR	0.39	0.08	1.89
EA-level % agree that family planning is for married women only	1.07	0.35	3.27
Cote d'Ivoire			
Reported covert use at P1 or P2	**0.41**	**0.21**	**0.80**
EA-level CPR	0.70	0.22	2.17
EA-level % agree that family planning is for married women only	1.03	0.49	2.15
India-Rajasthan			
Reported covert use at P1 or P2	**0.27**	**0.06**	**1.15**
EA-level CPR	0.17	0.01	2.06
EA-level % agree that family planning is for married women only	0.83	0.28	2.50

NOTE: Models also control for age, parity, wealth, education, residence (where available), pregnancies in P2 calendar, use episodes in the P2 calendar, and using a long-acting method in P1. Boldfaced odds ratios and CIs are significant at $p < 0.10$. Nigeria-Kano was excluded from the analysis of covert use due to small sample size, at $n < 100$. The survey question that defines covert use ("does your partner know you are using?") were asked of any woman who reported the use of any method except male condom, withdrawal, or male sterilization. The binary measure of covert use included in the model is 1=woman reported covert use at either P1 or P2, 0= woman never reported covert use. Therefore, this sample is limited to users at P1 or P2 of nonmale-controlled methods.

DISCUSSION

In this research, we used longitudinal panel data from nine geographies in seven countries to evaluate the reliability of the reproductive calendar. These geographies capture considerable variation in family planning characteristics, with a range of mCPR among all women from 7.6 percent to 48.5 percent and varying method mixes. To begin, we compared reports of contraceptive use, nonuse, and pregnancy from two separate datasets for the same women for

the same month. We evaluated the extent of agreement for three broad categories (nonuse, pregnancy, and contraceptive use) and 10 more-specific categories (nonuse, pregnancy, and contraceptive use for each specific method). We then identified characteristics associated with greater reliability in reporting these items, including both sociodemographic characteristics, measures of calendar complexity, and measures of the willingness of women to report contraceptive use.

Overall, across both the broader (3×3) and more specific approaches (10×10), we find that reliability generally falls within the "moderate to substantial agreement" category, with some results (Rajasthan) in the category of "excellent" reliability (Landis and Koch 1977). As expected, the reliability is generally higher for the broader approach, but the differences in Kappa statistics are not substantial between these approaches and fall between 0.00 and 0.02 for all geographies. Also, as expected, the extent of agreement increases with an expanded reporting period of +/− two months (Appendix Table 2), but there are still some remaining discrepancies with this buffer period, which suggests that there may be some items that are not reported at all on one calendar, as opposed to just inaccuracies in the reported timing of events.

Looking at results across geographies, reliability is not strongly associated with absolute levels of contraceptive use. For example, the Kappa statistics are about the same between Kenya, which has the second highest mCPR, and Nigeria-Kano, which has the lowest. Similarly, although Cote d'Ivoire and DRC-Kinshasa have about the same level of nonuse, the Kappa score is much lower for the latter (0.58 compared to 0.77).

However, reliability is associated with long-acting method use. When we limit our analysis to users, and to users of short-acting methods, the reliability declines from the analysis of all women. We find the lowest levels of agreement among women using short-acting modern and traditional methods. Only DRC-Kinshasa has a relatively high agreement among women using short-acting methods (at 75 percent), and most other locations have less than 60 percent agreement among these women. This suggests that the extent to which the calendar provides accurate data is influenced by the method mix, and the reliability of the calendar may improve in the future with increased use of long-acting methods.

Looking at individual-level characteristics, we find that there is an inverse relationship between the number of overall reproductive events reported in the calendar and the reliability of calendar reports: three other measures that are associated with less reliability in most countries include higher parity, a greater number of use episodes in the calendar, and not using a long-acting method at P1. In short, the more complex the calendar is, the less likely it is to be reliable. In contrast, sociodemographic measures are either not significantly associated with reliability, or not associated in most countries, which previous studies assessing the validity or reliability of survey reports have also found, including studies assessing the contraceptive calendar (e.g., Blanc et al. 2021; Callahan and Becker 2012; Carter et al. 2021; Strickler et al. 1997).

We find that reliability is associated with some of the measures that capture the woman's willingness to report contraceptive use as well. In six of the geographies where this was tested, we find that women who had not disclosed their contraceptive use to their husbands were less likely to be reliable in their reports than women whose husbands were aware of their contraceptive use. It is perhaps not surprising that women who did not report contraceptive use to

their husbands may also not consistently report this to an interviewer, which suggests that the tendency to reliably report contraceptive use may depend not only on the respondent's ability to remember their use patterns but also their willingness to report contraceptive use. We do not find consistent evidence for the other measures in this category, the EA-level average contraceptive use or the EA-level average percentage of women who believe family planning is only for married women; as these measures are only significantly associated with reliability in one or two settings. Nonetheless, the more consistent relationship between reliability and covert use suggests that reliability is partly influenced by women's willingness to report contraceptive use, and this influence should be further considered and explored.

Although we include a broader range of geographies in this analysis, some of our results are consistent with previous research. In the three previous studies, from urban Kenya, rural Bangladesh, and Morocco, Kappa statistics ranged from 0.56 to 0.79 across countries and the approaches to test reliability; these studies also generally found that women with more complex reproductive histories were less likely to reliably report contraceptive use (Callahan and Becker 2012; Strickler et al. 1997; Tumlinson and Curtis 2021). Unlike these earlier studies, however, we show the value of women's willingness to report contraceptive use, most prominently captured by covert use, as a significant predictor of reliability, and we show that reliability varies considerably across contexts in SSA.

As with other studies on this topic, an important limitation is that we cannot measure the validity of these reports, since we do not have a measure of the actual underlying status (pregnancy, contraceptive use, or nonuse). Although we believe that the cross-sectional reports of contraceptive use and pregnancy are more accurate than the retrospective calendar reports overall (reinforced by related analysis in Tsui et al. 2021), we have no way of knowing which reports are accurate, even when concordant. Other limitations include the inadvertent omission of two populations in our data collection: panel women who reported never having used family planning and who did not have an intervening pregnancy between P1 and P2, and panel women who were using family planning at the time of the P2 survey, who started within the calendar period, and who had no pregnancies during the calendar period (the latter occurred only in Kenya, Nigeria, DRC, and Burkina Faso). These omissions were only for a relatively small percentage of women in all countries, but these women may be systematically different in reliability than those included in our analysis.

Although we find that the overall quality of the contraceptive calendar data ranges from moderate to excellent, caution is necessary for some analyses of calendar data. As described above, calendar complexity (higher parity, more use of contraceptive episodes, and use of short-acting methods) is associated with lower reliability, which has important implications for analyses of contraceptive dynamics and change, such as contraceptive discontinuation and switching, as such behaviors are less likely to be measured accurately in the calendar.

What can be done to improve the reliability of reports of contraceptive use overall and the calendar in particular? Based on this research, we offer several observations. First, we emphasize the importance of paper aid. Given the complexity of the calendar, it would be impossible to effectively view the completed calendar on the mobile phone used for data collection, so the paper aid is used to view the full reproductive calendar for coherence before it is entered on the phone. While there is the potential for data entry error of the paper aid, PMA took pictures of the aid after data collection to (1) compare with the entered data after

data collection (and we found minimal data entry errors), and (2) provide the potential for correcting the data if necessary. An experiment in which REs were randomly assigned to use the paper aid compared with no aid would be necessary to identify the effect of the aid on calendar reliability. Second, because prior research has suggested that the reliability of calendar data decreases with larger recall periods (Bradley et al. 2015), PMA chose to implement a two-to-three-year calendar; and it is very likely that data quality would be worse with a five-year calendar instead (as DHS and other surveys do). Third, PMA devoted a considerable amount of time training REs in the calendar data collection, including extensive pilot testing of the approach prior to data collection, and video instructions that REs could keep on their phones for future reference. Fourth, previous analysis of PMA data has demonstrated the value of the RE approach, which suggests that REs yield better quality data than using interviewers who are not from the study sites (Anglewicz et al. 2019; Safi 2019). Based on this analysis, it is reasonable to expect that the impact of social desirability bias might be greater with "outsider" interviewers. Finally, because we find that the use of long-acting methods is associated with greater reliability, the increase in these methods in recent years (Tsui et al. 2017) suggests that calendar data will be more reliable over time, although one would also want to consider the fertility rate and extent of method switching when using the calendar approach.

Finally, we revisit the tradeoffs between study designs. The longitudinal panel approach allows a comparison of reports for the same women over time, which permits one to identify characteristics associated with reporting patterns; as well as the opportunity prospectively measure contraceptive use. In contrast, a cross-sectional design only allows a population-level comparison but is less costly than a longitudinal panel. If evaluating the reliability of the calendar data is a goal of the study, the longitudinal approach is preferable.

This research was primarily focused on measuring the reliability of the contraceptive calendar and the factors associated with consistent reports. We, therefore, contribute to the discussion of how to best measure reproductive histories in surveys, but more discussion is needed. Critical questions remain such as: do misreports of contraceptive use cause significant errors in key measures like contraceptive discontinuation rates, or are these errors minor and tolerable?

DATA AVAILABILITY STATEMENT

Data for this analysis can be accessed both from the PMA website (at https://www.pmadata.org/data/available-datasets) and IPUMS (https://pma.ipums.org/pma/). These data can be accessed by anyone after submitting a short request form.

ETHICAL APPROVAL AND PATIENT-INFORMED CONSENT

Ethical approval for the study protocol was provided by Johns Hopkins University and the appropriate ethical body within each country. All study participants were enrolled via an informed consent protocol.

FUNDING

REFERENCES

Ali, Mohamed M., John Cleland, and Iqbal H. Shah. 2012. *Causes and Consequences of Contraceptive Discontinuation: Evidence from 60 Demographic and Health Surveys*. Geneva: World Health Organization.

Amin, Ruhul, Nirali Shah, Becker Stan. 2010. "Socioeconomic Factors Differentiating Maternal and Child Health-Seeking Behavior in Rural Bangladesh: A Cross-Sectional Analysis." *International Journal for Equity in Health* 9(1): 1–11.

Anglewicz, Philip, Pierre Akilimali, Linnea Perry Eitmann, Julie Hernandez, and Patrick Kayembe. 2019. "The Relationship Between Interviewer–Respondent Familiarity and Family Planning Outcomes in the Democratic Republic of Congo: A Repeat Cross-Sectional Analysis." *BMJ Open* 9(1): e023069.

Becker, Stan, and Nafissatou Diop-Sidibe. 2003. "Does Use of the Calendar in Surveys Reduce Heaping?" *Studies in Family Planning* 34(2): 127–132.

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* 23(6):386–339.

Bignami-VanAssche, Simona. 2003. "Are We Measuring What We Want to Measure? Individual Consistency in Survey Response in Rural Malawi." *Demographic Research* S1: 77–108.

Blanc, Ann K., Katharine J. McCarthy, Charlotte Warren, Ashish Bajracharya, and Benjamin Bellows. 2021. "The Validity of Women's Reports of Family Planning Service Quality in Cambodia and Kenya." *Studies in Family Planning* 52(1): 77–93.

Bradley, Sarah E. K., William Winfrey, and Trevor N. Croft. 2015. "*Contraceptive Use and Perinatal Mortality in the DHS: An Assessment of the Quality and Consistency of Calendars and Histories*." DHS Methodological Reports No. 17. Rockville, MD: ICF International.

Callahan, Rebecca L., and Stan Becker. 2012. "The Reliability of Calendar Data for Reporting Contraceptive Use: Evidence from Rural Bangladesh." *Studies in Family Planning* 43(3): 213–222.

Carter, Emily, Karen Chang, Luke Mullany, Subarna Khatry, Steven LeClerq, Melinda Munos, and Joanne Katz. 2021. "Reliability of Maternal Recall of Delivery and Immediate Newborn Care Indicators in Sarlahi, Nepal." *BMC Pregnancy and Childbirth* 21(1): 1–13.

Curtis, Sian L., and Ann Blanc. 1997. "*Determinants of Contraceptive Failure, Switching, and Discontinuation: An Analysis of DHS Contraceptive Histories*." DHS Analytical Reports No. 6. Calverton, MD: Macro International Inc.

FP2020. 2021. "FP2020 Core Indicators." Available at: http://www.track20.org/pages/data_analysis/core_indicators/overview.php. Accessed May 17, 2021.

Goldman, Noreen, Lorenzo Moreno, and Charles F. Westoff. 1989. "Collection of Survey Data on Contraception: An Evaluation of an Experiment in Peru." *Studies in Family Planning* 20(3): 147–157.

Knodel, John, and Sauvaluck Piampiti. 1977. "Response Reliability in a Longitudinal Survey in Thailand." *Studies in Family Planning* 18(9): 55–66.

Landis, J. Richard, and Koch, Gary G. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33(1): 159–174.

Performance Monitoring for Action. 2022. "Survey Results Summaries, Phase 1 Data Collection." Available at: https://www.pmadata.org/data/survey-results-summaries. Accessed October 27, 2022.

Polis, Chelsea, Sarah Bradley, Akinrinola Bankole, Tsuyoshi Onda, Trevor Croft, and Susheela Singh. 2016. *Contraceptive Failure Rates in the Developing World: An Analysis of Demographic and Health Survey Data in 43 Countries*. New York: Guttmacher Institute.

Safi, Sally. 2019. "*The Effect of Innovative Survey Design on Demographic and Reproductive Health Data in Sub-Saharan Africa: A Multi-Country Analysis.*" Doctoral dissertation, Johns Hopkins University.

Strickler, Jennifer, Robert Magnani, Gilman McCann, Lisanne Brown, and Janet Rice. 1997. "The Reliability of Reporting of Contraceptive Behavior in DHS Calendar Data: Evidence from Morocco." *Studies in Family Planning* 28(1): 44–53.

Tsui, Amy, Win Brown, Qingfeng Li. 2017. "Contraceptive Practice in Sub-Saharan Africa." *Population and Development Review*, 43(S1): 166.

Tsui, Amy O., Carolina Cardona, Varsha Srivatsan, Funmilola OlaOlorun, Elizabeth Omoluabi, Pierre Akilimali, Peter Gichangi, Mary Thiongo, Scott Radloff, Philip Anglewicz, and The PMA Agile Team. 2021. "Is Client Reporting on Contraceptive Use Always Accurate? Measuring Consistency and Change with a Multicountry Study." *Studies in Family Planning* 52(3): 361–382.

Tumlinson, Katherine, and Siân Curtis. 2021. "Assessing the Reliability of the Retrospective Reproductive Calendar: Evidence from Urban Kenya." *Studies in Family Planning* 52(4): 467–486.

United Nations Department of Economic and Social Affairs, Population Division. 2020. "World Fertility and Family Planning 2020: Highlights." ST/ESA/SER.A/440.

Westoff, Charles, Noreen Goldman, and Lorenzo Moreno. 1990. "*Dominican Republic Experimental Study: An Evaluation of Fertility and Child Health Information.*" Office of Population Research, Princeton University.

Zimmerman, Linnea, Hannah Olson, Amy Tsui, Scott Radloff, and PMA2020 Principal Investigators Group. 2017. "PMA2020: Rapid Turn-Around Survey Data to Monitor Family Planning Service and Practice in Ten Countries." *Studies in Family Planning* 48(3): 293–303.

ACKNOWLEDGMENTS