Endogenous Sample Selection*

Torsten Figueiredo Walter[§] Niclas Moneke[¶]

April 23, 2026

This paper examines how incentives for data collectors shape the selection of survey samples. We provide causal evidence that data collectors manipulate samples by screening out subjects that require high effort. Exploiting random variation in the number of questions to be asked about individuals across 3.4 million households in 181 surveys worldwide, we find evidence of such endogenous sample selection in 72% of surveys. On average, 8.8% of individuals eligible for extensive interviewing are missing in these surveys. Selection out of sample is systematic: missing individuals disproportionately come from marginalised populations. As a result, policy-relevant aggregate statistics are biased. We estimate fertility to be 5% lower, child mortality 1% higher and child marriage 7% lower than reported. Finally, we show that sample selection matters for research. Prominent causes of economic development, such as climate and institutions, also cause differential sample selection, providing a cautionary tale for identification.

Keywords: Data collection, moral hazard, sampling, selection
JEL classification: C81, C83, J22, O11, O12

*We thank Alison Andrew, Oriana Bandiera, Gharad Bryan, Raj Chetty, Andrew Dillon, Kevin Donovan, Dennis Egger, Ben Faber, Luis Garicano, Maitreesh Ghatak, Doug Gollin, Sydney Gourlay, Stephane Helleringer, Lisa Ho, Talip Kilic, Gabriel Kreindler, Ted Miguel, Rohini Pande, Imran Rasul, Mark Rosenzweig, Martin Rotemberg, Nick Ryan, Chris Udry, Chris Woodruff and various seminar and conference audiences at Berkeley, BU, BREAD Lahore, CEPR Paris, CSAE Annual, Durham, ESEM Rotterdam, IFS-UCL, IPA, Kent, NBER SI, NEUDC, Northwestern, Notre Dame, Nova SBE, NYU, NYU Abu Dhabi, Oxford, PacDev, Philadelphia Fed, PSE, Reading, Sussex, STEG Annual, UAB, UIUC, UVA, USC, WUSTL and Yale for helpful comments. Claudia Alcarra, Ana Barros, Frederic Cochinard, Muskaan Chopra, Lakshmipriya Gopikrishnan, Ameni Hajji, Abigail Lamptey, Sophia Lin, Shifa Muhammad, Alfred Shauri, Jordan Simpson, Kristina Sisiakova, Kevishen Valeyatheepillay, Emma Wang and Ruike Zhang provided excellent research assistance. Financial support from STEG is gratefully acknowledged.

[§]New York University Abu Dhabi, Office A5-193, Abu Dhabi, UAE. Email: t.f.walter@nyu.edu
[¶]Department of Economics, University of Oxford, UK. Email: niclas.moneke@economics.ox.ac.uk

1 Introduction

Human-collected data are ubiquitous. Doctors gather medical information, police officers record crime cases, and field interviewers collect survey responses. All of these data collectors have an incentive to avoid collecting information that is particularly costly to obtain. They also frequently have the means to do so by excluding observations from the record that require disproportionate effort. If exclusion is correlated with the potential information collected, then this *endogenous* selection of the sample introduces systematic bias and can lead to erroneous inference – which is of particular concern given the essential role of human-collected data for policy and research.

In this paper, we examine endogenous sample selection in surveys, and ask three questions: first, do data collectors systematically exclude high-cost subjects? Second, is the resulting selection of subjects out of the sample non-random? Third, does this selection affect inference and analysis, and if so, how?

In theory, sampling for surveys is fully determined by protocol, but in practice, data collectors exert considerable influence over sampling as they commonly fulfill two separate tasks: first, they screen for eligible subjects among a target population using pre-established eligibility criteria. This screening determines which subjects are included in the sample. Second, they collect detailed data about the sampled subjects. Hence, successful screening of subjects creates more work for data collectors, introducing an incentive to sabotage the screening, thereby manipulating the sample.

Consider a simple example for illustration. Household surveys, such as the Demographic and Health Survey (DHS) or the Multiple Indicator Cluster Survey (MICS), primarily collect information about children (aged 0-5) and women (aged 15-49) by administering long individual questionnaires about them. To identify eligible women and children in the first place, data collectors list all household members and screen for eligibility based on sex and age. This setup creates an incentive for data collectors to either manipulate members' eligibility criteria or omit eligible members entirely.

Figure 1 illustrates how these dynamics play out in the 2006 MICS from Togo: The top panels show that question load for eligible women and children of either sex is about three times as high as for ineligible household members. The bottom panels highlight how the associated age distributions lack mass in all age ranges that are eligible for individual questionnaires (grey-shaded areas) and have excess mass on the ineligible side of eligibility thresholds. Reassuringly, the male age distribution (bottom right) shows the same missing mass below the age of 5 as the female age distribution, but does not display missing mass between 15 and 49 (gold-shaded area), thereby suggesting a causal link between question load and sample inclusion.

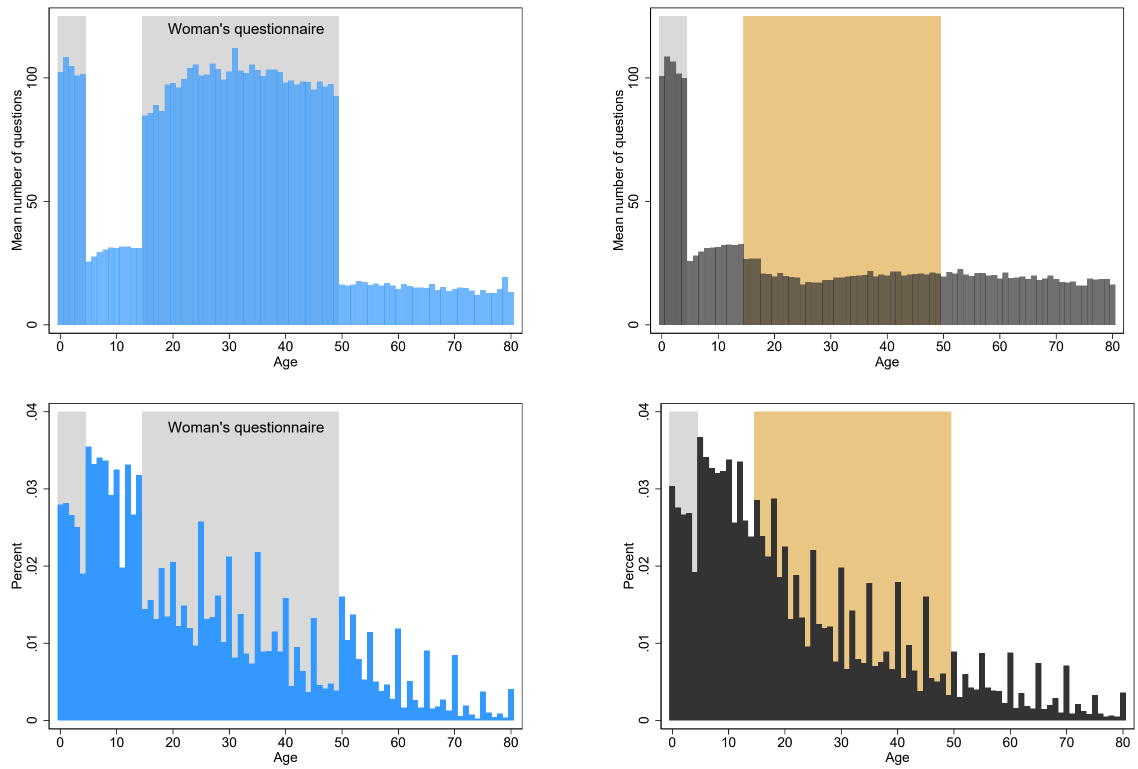


Figure 1: Question load (top) and age distribution (btm) by sex in Togo MICS 2006

This figure plots the mean number of questions asked about female and male household members by age in panels (a) and (b), and their age distributions in panels (c) and (d). Age groups eligible for individual questionnaires are shaded in grey. Children under the age of 5 are eligible for the 'under-five questionnaire' and women between the ages of 15 and 49 are eligible for the 'woman's questionnaire'. In the right-hand panels, the age range between 15 and 49 is shaded in gold to facilitate comparison with the same range in the left-hand panels.

This paper provides global evidence that variation in data collector effort cost across survey subjects leads to endogenous sample selection, thereby biasing aggregate statistics and undermining causal identification in empirical research. First, we exploit random assignment of question load across 3.4 million households across 181 surveys worldwide to show that data collectors manipulate survey samples by excluding high-cost subjects. Second, we demonstrate that manipulation leads to non-random selection of subjects out of sample, resulting in under-representation of marginalised populations. Third, we highlight that this selection introduces bias in policy-relevant statistics and undermines causal identification in empirical research.

Throughout our analysis, we view data collectors as economic agents collecting data on behalf of their principals, such as a national statistical offices or teams of researchers. The data collectors' actions are imperfectly observable. In their decision to truthfully record a survey subject, e.g., a household member, as eligible or not for subsequent data collection, they weigh the effort cost of additional data collection against the expected penalty for manipulation. Hence, all else equal, they are more likely to exclude subjects with higher effort cost and lower expected penalty.

To investigate endogenous sample selection empirically, we leverage data from two of the largest international household survey programs – the Demographic and Health Survey (DHS) and the Multiple Indicator Cluster Survey (MICS). These surveys are widely used in economics and beyond. Since 2010, at least 37 papers in top general interest economics journals have made use of these survey data. Their use outside of economics is even greater. In 2022 alone, close to 1,000 articles in journals included in the *Web of Science Essential Science Indicators* across the natural sciences, social sciences and the fields of health and medicine referred to at least one of the two survey programs in their title or abstract. At the same time, the DHS and MICS are of great importance for policy, e.g., to monitor the Sustainable Development Goals, to allocate aid, or to inform the design of national policies.

To estimate the causal effect of data collector effort cost on sample exclusion, our main empirical strategy exploits the random assignment of individual questionnaires for men ('man's questionnaire') across households in 181 DHS and MICS. In this context, data collectors typically work on temporary contracts for the duration of each survey and receive a fixed daily wage. Administering individual questionnaires is time-consuming, with each additional man's questionnaire requiring on average 25 minutes of extra work. Disutility of effort and pressure to adhere to pre-determined survey schedules create an incentive for data collectors to shorten household interviews by reducing the number of household members eligible for individual questionnaires. Indeed, we find that in 72% of surveys, the number of men eligible for the man's questionnaire is significantly smaller in households that have randomly been chosen to receive the man's questionnaire (henceforth referred to as treatment households). The average reduction in these surveys amounts to 8.8% of eligible men.

Since many policy-relevant statistics, such as fertility and child mortality, are derived from the individual questionnaire for women ('woman's questionnaire'), it is important to understand to which extent endogenous sample selection also affects eligible women. In the absence of random assignment of the woman's questionnaire across households, we fall back on an alternative empirical strategy that compares the number of female household members of eligible and ineligible age in DHS and MICS to those in contemporaneous population censuses. Whereas the number of questions to be administered to women of eligible age (typically 15-49 years) is much larger than the number of questions to be administered to women outside this age range in DHS and MICS, no such difference in question load between women of eligible and ineligible age exists in population censuses. Therefore, data collectors face strong incentives to exclude women of eligible age in DHS and MICS, but have no such incentive in census data collection. Leveraging 75 survey-census pairs from 38 countries, we find strong evidence of endogenous sample selection mirroring our findings for men: we estimate that at least 91% of surveys record less eligible women than contemporaneous censuses, with an average relative lack of eligible women of 6.7%.

Our simple theoretical framework provides us with two testable empirical predictions. First, that exclusion of household members from the sample is increasing in data collector effort cost. Second, that it is decreasing in the probability of detection of sample manipulation. We provide evidence in support of both predictions. Exploiting exogenous variation in day-to-day temperature within data collector, we estimate the causal effect of higher effort cost from collecting data under extreme temperatures on sample manipulation. We find an U-shaped relationship, with wet-bulb temperatures above $25°C$ and below $10°C$ increasing the number of eligible men missing in treatment households. In addition, we provide suggestive evidence on the impact of systematic backchecks. Comparing surveys with and without such checks, we find that the average share of missing eligible men amounts to 2.6% in the former as opposed to 8.0% in the latter.

If women and men of eligible age were excluded from samples at random, our findings would be of minimal concern to data users. However, we show that missing individuals are not randomly selected by data collectors and are best described as belonging to marginalised populations. By comparing eligible men in treatment and control households, we find that missing men are less closely related to the head of their respective households, less likely to have ever been married, less educated, less healthy and poorer. A comparison of the characteristics of women of eligible age in the DHS/MICS and contemporaneous censuses yields comparable conclusions for missing women. Taken together, our findings on selection suggest that among the high-effort-cost household members – men and women of prime age who are eligible for individual questionnaires – data collectors screen out exactly those individuals at the margins of their respective households where household definitions and cultural norms leave room for interpretation, and chances that supervisors can detect manipulation are lower. Put differently, the nuclear family, i.e., household members whose absence is easiest to detect, are relatively over-represented in the data.

How does the relative under-representation of marginalised populations affect aggregate statistics? Using both of the above empirical strategies, we estimate that one of the original core outcomes of the DHS, fertility, is upward-biased by approximately 5% due to endogenous sample selection.[1] This upward bias matters: in the absence of alternative data sources, population projections and public resource allocation in many low-income countries directly depend on fertility estimates from the DHS. In fact, the former are highly sensitive to even moderate biases in fertility and the discrete nature of resource allocation makes the latter highly sensitive to small changes in population, as demonstrated by Figueiredo Walter et al. (forthcoming).

The DHS and MICS also provide crucial input for the monitoring of the United Nations' Sustainable Development Goals (SDGs). To assess the quantitative importance of endogenous sample selection in the absence of alternative data sources on the SDGs, we correct survey samples for selection on *observables*. Our findings suggest that reported child mortality is understated by 1% and child marriage overstated by 7% on average. In 25% of surveys, these deviations exceed 2% and 12%, respectively. Other SDG indicators derived from the DHS/MICS appear similarly biased.

Does endogenous sample selection also have implications for empirical research? Endogenous sample selection poses a threat to the causal identification of treatment effects if it is correlated with treatment. To assess the relevance of this concern, we study the effect of prominent causes of economic and social development on endogenous sample selection. We focus on three broad sets of causes – climate shocks, institutions and economic origins – and leverage empirical strategies from eight existing studies for identification. We find that all of the considered drivers of development also affect endogenous sample selection, thus cautioning against the causal interpretation of associations between these drivers and outcome data collected by humans in the field.

[1] Note that the predecessor of the DHS in the 1970s and 1980s was the World *Fertility* Survey.

Overall, our findings highlight a fundamental trade-off in data collection between the amount of information collected from respondents and bias in the sample of respondents. In our set of surveys, we estimate an average elasticity of sample size with respect to question load of -0.01, suggesting that adding an individual questionnaire that includes the same number of questions as the household roster to a survey leads to a reduction in eligible survey subjects by 1%. Beyond this set of surveys, we uncover evidence consistent with endogenous sample selection in a wide array of commonly used data products in low- and high-income countries alike, confirming the widespread nature of the phenomenon and the resulting information-bias trade-off.

This paper contributes to three streams of literature. First, it adds to a long and active literature on selection in surveys (Rubin, 1976; Meyer et al., 2015; Dutz et al., 2021). While this literature is largely focused on non-response bias, i.e., self-selection of respondents, this paper highlights an overlooked margin of selection that, unlike non-response, is not (directly) observable to the econometrician: the selection of respondents by data collectors. We show that endogenous sample selection driven by data collector incentives is quantitatively of similar importance as non-response and can likewise lead to substantial bias in aggregate statistics and threaten causal identification in research.

Second, we contribute to an emerging literature on the economics of data collection that analyses how incentives affect the behaviour of data collectors. On the one hand, recent work studies classic questions of organisational economics in the context of survey production, including the effects of pay (Menold et al., 2018; Kim et al., 2020) and supervision arrangements (Sen, 2024). We examine the incentives for data collectors that arise from survey design and the resulting distribution of effort cost across respondents, similar to Figueiredo Walter et al. (forthcoming). On the other hand, cross-disciplinary work on data fabrication implicitly attributes the fabrication to incentives faced by data collectors (Crespi, 1945; Kosyakova et al., 2014; Blasius & Thiessen, 2015; Finn & Ranchhod, 2015). While the manipulation of respondent screening by data collectors has been recognised as a potential concern by practitioners (Marckwardt & Rutstein, 1996; Pullum, 2006), it has barely received any attention from researchers.[2] We provide new insights on how the allocation of question load across survey subjects translates into incentives for data collectors to falsify survey

[2] One exception is Eckman and Koch (2019) who study differences in data collector involvement in the screening of respondents across countries in the European Social Survey.

responses with regards to eligibility criteria.

Third, this paper adds to a broad literature on the design and implementation of data collection. Within this literature, our work is most closely related to research on enumerator effects (West & Blom, 2017; Di Maio & Fiala, 2019). While we also examine the role of data collectors, we do not focus on the effect of the identity of data collectors, but investigate how variation in incentives (within data collector) affects data collection. In addition, we relate to recent work on respondent fatigue (Ambler et al., 2021; Abay et al., 2022; Jeong et al., 2023) that examines the effect of question load on respondent behaviour. In this paper, we instead shed light on the effect of question load on the behaviour of data collectors.

The remainder of the paper is structured as follows. Section 2 introduces a simple theoretical framework. Section 3 provides background information on the DHS and MICS. Sections 4 and 5 present our empirical strategies and our main results on missing household members and their selection out of sample. Section 6 outlines implications of endogenous sample selection for policy and research, followed by Section 7 which elaborates on the external validity of our findings. Finally, Section 8 discusses the information-bias trade-off at the heart of endogenous sample selection and its relevance for data collection more generally, before Section 9 concludes.

2 Theoretical framework

A simple theoretical framework of data collector behaviour can guide the empirical analysis of endogenous sample selection. Data collectors face a choice of including or excluding a given subject from the eligible population in the initial listing exercise.[3]

The decision to report or not to report a given subject, $R \in \{0, 1\}$, in the listing is determined by two competing forces: the cost c of enumerating a given subject if the data collector reports them, and the probability of detection p (and associated penalty of losing their wage w) if they do not report. Both forces vary with the subject's observable characteristics x, such as a their age or sex.

The utility of the data collector, U, as a function of the decision to report, R, can be written as follows:

[3] This decision to exclude covers both empirically relevant cases we document in Section 4: displacement, i.e., adjusting a subject's age to make them ineligible, and omission, i.e., not reporting a subject on the roster. Both cases lead to exclusion of that subject from the long questionnaire.

$$U(R) = (1 - [p(x) \times (1 - R)])w - c(x) \times R \tag{1}$$

The data collector will decide to report a given subject $R = 1$ if their wage w exceeds the detection probability-scaled cost of enumeration:

$$U(1) > U(0) \quad if \quad w > \frac{c(x)}{p(x)} \tag{2}$$

Therefore, a given subject is more likely to not be reported if their enumeration cost is high, or the detection probability of them not having been reported is low. As illustrated in Figure 1 for household surveys with individual-level questionnaires, the cost of enumeration, driven by question load, varies across household members as a function of their observable characteristics, vector x, e.g., age and sex. For example, a 16-year old female is eligible for a long questionnaire in the DHS/MICS surveys, incurring higher cost for data collectors than a 14-year old female or 16-year old male.

However, deciding not to include a subject entails a risk of detection by the supervisor. With probability p manipulation in the form of not reporting a subject is detected, the data collector gets fired and does not receive their wage w. Detection is not random, and we assume the detection probability to depend on observable characteristics x. In our setting, some observables that aid detection also drive data collector cost, such as a household member's age, whereas other observables are independent of cost: the genealogical distance to the household head, ceteris paribus, does not affect data collector cost – but a household without a wife or husband is much easier to detect for supervisors than one missing an aunt, cousin, niece or grandchild.

The main testable prediction arising from our theoretical framework is that data collectors maximise utility of data collection by not reporting, i.e., excluding, high cost, low detection probability subjects. In Section 4 we test if data collectors exclude household members from the sample in general. Exogenous variation in data collector effort cost (or variation in detection probability) allows us to explicitly test the mechanism described above in Section 4.3. If data collectors exclude household members as a function of subjects' expected effort cost and perceived detection probability, this can introduce systematic selection of the sample, which we test for in Section 5.

3 Background

3.1 Relevance

In this paper, we primarily study endogenous sample selection in two large international household survey programs, the Demographic and Health Survey (DHS) and the Multiple Indicator Cluster Survey (MICS). The DHS focuses on fertility, family planning, maternal and child health, HIV/AIDS, malaria, and nutrition. It was funded by USAID and implemented by ICF. Since its inception in 1984, the program conducted more than 400 surveys, with sample sizes ranging from around 2,000 households in the early years to more than 40,000 more recently and an estimated average cost of USD 1.6 million per survey. The MICS program, funded by UNICEF, bears many similarities with the DHS. It also mainly focuses on the situation of children and women in low- and middle-income countries and comprises more than 350 surveys. Sample sizes and costs tend to be lower, however, averaging around 12,000 households and USD 1.1 million per survey, respectively. Both survey programs have a reputation for collecting accurate, comparable, nationally representative data using standardised, state-of-the-art survey instruments across countries.[4]

We focus on these two household survey programs for three reasons. First, they are of great importance for research, especially in the social sciences and the fields of medicine and health.[5] As data from the Web of Science database shows (Figure A1), the annual number of articles published in reputable journals across all fields that refer to the DHS or MICS in their title or abstract has increased 27-fold since 2000, reaching nearly 1,000 in 2022.[6] The true use of the data is likely much higher, though, because many papers use the data without explicit mention in title or abstract. For example, out of 37 papers published in top general interest journals in economics since 2010 that use the DHS/MICS, only one refers to them in title or abstract.

Second, the DHS and the MICS are of great importance for policy. They are key to monitoring the Sustainable Development Goals (SDGs), providing input data for about 30 SDG indicators. They affect aid flows, not least through programs that are explicitly conditioned on DHS-derived indicators, such as the World Bank Program-

[4] See Sustainable Development Solutions Network (2015) for details on survey cost estimates.

[5] Short Fabic et al. (2012) provide a historic overview of DHS use in population/health research.

[6] Statistics based on Web of Science database keyword search, restricted to journals that formed part of the Essential Science Indicator journal master list as of June 2024.

for-Results. At the national level, they are an important input to policy, in particular in the health sector, as documented by Nolan et al. (2017) and evidenced by frequent references to them in national health policy plans.

Third, both survey programs have *global* coverage. Since program inception, the DHS and the MICS program have conducted surveys in more than 90 and 120 countries, respectively, making them a unique source of globally comparable data over a time span of more than 30 years.[7]

3.2 Survey design

USAID/ICF and UNICEF provide questionnaire templates to local statistical agencies at the beginning of each survey wave. The DHS originally consisted of two questionnaires: a household questionnaire (including household roster) and a woman's questionnaire. The MICS was originally composed of three questionnaires: a household questionnaire (including household roster), a woman's questionnaire and an under-five questionnaire. In both survey programs, the household questionnaire is composed of two parts, the household roster and household-level questions. The household roster gathers basic demographic information on all household members and is used to determine the eligibility of household members for individual questionnaires based on sex and age. Household-level questions concern topics such as asset ownership, energy use and sanitation. The woman's questionnaire is administered to all women aged 15 to 49 and focuses on fertility and maternal health; the under-five questionnaire to all children under 5, focusing on child health and development.

In later survey phases, both survey programs introduced a man's questionnaire. This questionnaire addresses similar topics as the woman's questionnaire – mainly fertility, health and sexual behaviour – but is typically much shorter. In most surveys, the eligible age ranges from 15 to 49, but in some cases it also includes older men up to the age of 54, 59 or 64. Importantly, in many surveys this questionnaire is only administered in a random subset of households within each enumeration area.

Individual questionnaires are administered after the household roster has been completed. This implies that at the time of the roster completion, the survey respondent does not know how the age and sex of household members recorded in the roster affect the length of the household interview. Data collectors are very much aware of

[7]Statistics retrieved from the official DHS website – https://dhsprogram.com/ – and the official MICS website – https://mics.unicef.org/ – on August 18, 2024.

this, however, since they are familiar with the survey structure from their training and their experience with previous households. Moreover, the survey instruments make the eligibility of household members for individual questionnaires very salient to data collectors, asking them to explicitly mark every eligible member as they fill in the roster (see Figures A2 and A3 for illustration).

An important difference between the DHS and the MICS lies in the household definition they work with. The MICS operates with a *de jure* household definition, recording all usual members. Each of these members qualifies for the individual questionnaire if they are in the eligible age range. The DHS instead records all usual household members *and* all guests who stayed in the household the night before.

The eligibility of *de facto* and *de jure* household members for individual questionnaires, however, varies across surveys. In phases 1 and 2 of the DHS program, eligibility was conditional on having slept in the household last night. From phase 3 onwards, most surveys did not condition eligibility on having slept in the household last night anymore. However, all results published by the DHS remain restricted to *de facto* populations to avoid double-counting.[8] Therefore, we define eligibility in the DHS as being of eligible age and having slept in the household last night throughout.

3.3 Data collector incentives

DHS and MICS are funded and supported by USAID and UNICEF, respectively. Both programs provide questionnaire templates that are standardised within survey phases and guidelines for implementation in the form of manuals for data collectors, supervisors, editors as well as data collector training, household sampling and other topics. However, surveys are ultimately implemented by local agencies, most commonly National Statistical Offices.[9] Hence, data collectors are recruited locally. Nonetheless, hiring practices barely vary across contexts. Temporary contracts for the duration of the survey are standard. Only a few implementing agencies rely on their permanent staff for enumeration in addition to temporary workers.[10] Data collectors generally have to meet the following criteria: They have to (i) be available to

[8] In fact, none of the data from individual interviews of household members who did not sleep in the household last night is published.

[9] 82% of the surveys in our main sample were implemented by National Statistical Offices, 15% by other governmental bodies, such as Ministries of Health, and 3% by nongovernmental organisations.

[10] Fieldworker data from recent DHS confirm that most data collectors work under temporary contracts. In the 19 surveys included in our main sample for which fieldworker data is available, on average 13% of data collectors are permanent employees and 87% have temporary contracts.

work full time for the duration of the survey, (ii) exceed a minimum level of physical fitness, so they can walk long distances, and (iii) speak at least one of the languages used for training. Additionally, there is a preference for local candidates from within a region of a country and candidates with secondary or higher education. As a result, interviewers are more educated than the average respondent in most contexts.

Data are collected by enumeration teams usually comprised of a supervisor, a field editor and several data collectors. Supervisors are in charge of the organisation of the fieldwork, including the assignment of households and questionnaires to data collectors and spot check re-interviews. Field editors are responsible for monitoring data quality. To this end, they observe interviews, edit completed questionnaires and may ask data collectors to return to interviewed households to correct problems. Additional data quality issues can be detected through field check tables produced by data processing teams during fieldwork. These are typically provided to supervisors after the completion of an enumeration area and can inform measures to improve data quality going forward. All of this implies that the missing eligible individuals we detect in this paper were either not flagged in any of the data quality checks or, if flagged, they were not addressed successfully.[11]

Data collectors' employment contracts are designed by the implementing agencies. Thus, they can vary across surveys. In practice, however, data collectors are almost always paid a fixed daily wage plus a per diem for food and accommodation. The daily workload of enumeration teams is typically set in advance by the central office of the implementing agency and adherence to the schedule is heavily emphasized during fieldwork. Supervisors are responsible for assigning households to data collectors at the beginning of each day, but these assignments can be adjusted throughout the day as some interviews take shorter or longer than expected. Data collector performance is monitored continuously throughout the survey. Supervisors complete a so-called 'interviewer progress sheet' after the completion of each survey cluster to track how data collectors are keeping up with the assigned workloads.[12] This means that data collectors benefit from missing eligible household members in at least two ways. First, they will be better able to keep up with the assigned workloads, thereby building a good reputation, minimising their risk of termination, and increasing their chances of

[11] Neither in the DHS nor the MICS data is it possible to observe which interviews were monitored by a field editor or re-conducted by a supervisor.

[12] See LoPalo's (2023) Online Appendix Figure 1 for the DHS 'interviewer progress sheet'.

re-employment.[13] Second, they may have shorter working days.

The incorrect completion of household rosters also carries a risk for data collectors. Supervisor guidelines indicate that terminations may be necessary if data is falsified. It is unclear how common such terminations are in practice, but the DHS recommends to recruit reserve data collectors who can step in after separations.[14]

4 Missing household members

4.1 Missing men

Empirical strategy. In many DHS and MICS, only a random subset of households is eligible for the man's questionnaire. Random assignment is carried out at the headquarters of the implementing agency after household listing in all enumeration areas has been completed.[15] To this end, USAID and UNICEF provide implementing partners with a computer tool for randomisation.[16] Each household's randomly drawn eligibility status is pre-filled on their questionnaire and thus visible to data collectors.

Relying on the random assignment of the man's questionnaire across households, we run the following OLS regression to estimate the causal effect of eligibility for the man's questionnaire:

$$Y_{ic} = \alpha_c + \beta MQ_{ic} + \epsilon_{ic} \tag{3}$$

where Y_{ic} is an outcome of interest of household i in stratum c. MQ_{ic} is an indicator variable that takes the value one if household i is eligible for the man's questionnaire, and zero otherwise. α_c is a set of stratum fixed effects. In most surveys, strata correspond to enumeration areas.[17] The regression coefficient β captures the causal effect of household assignment to the man's questionnaire on the outcome of interest.[18]

We attribute the difference in outcomes between treatment and control households to the difference in incentives faced by the data collector. While we concede that the

[13] DHS fieldworker data shows that many data collectors have previous DHS experience.

[14] This subsection is based on exchanges with UNICEF's Data Collection Unit, and LoPalo (2023).

[15] Household listing in all survey clusters is typically conducted months ahead of survey fieldwork.

[16] The MICS sampling and randomisation tool is available here.

[17] In a few MICS, the random assignment of the man's questionnaire is additionally stratified by the presence of children below age 5, as recorded during the preceding household listing exercise.

[18] Note that we do not observe any cases of eligible men in eligible households that were not attempted to be interviewed individually. Hence, data collectors appear to comply perfectly with the random assignment. But not all eligible men complete the individual interview. Some are absent, incapacitated or refuse at any point. The average completion rate in our sample of surveys is 90.5%.

listing of household members is ultimately the product of the interaction between data collector and respondent – and respondents may also be interested in shortening the interview – we argue that the respondent does not possess the necessary information that would allow them to intentionally reduce the number of eligible men. First, respondents are unlikely to be familiar with the question load distribution across age since neither the DHS nor the MICS are panel surveys, but repeated cross-sections that are typically only carried out every 5 years.[19] Second, unlike data collectors, respondents do not know the eligibility of their household for the man's questionnaire.

Data. Based on the universe of survey reports published on their official websites, we identify 181 surveys, 135 DHS and 46 MICS, carried out across 73 countries between 1991 and 2022 in which a man's questionnaire was administered to a random subset of households. Figure 2 illustrates their geographic coverage, including low- and middle-income countries from all continents (see Table A2 for a complete list). The resulting dataset includes 3.4 million households out of which 1.1 million were randomly assigned a man's questionnaire.[20]

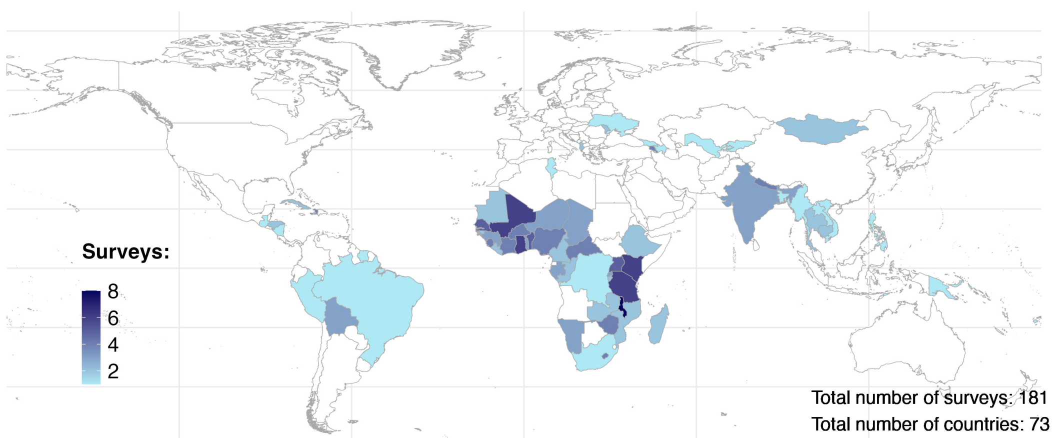


Figure 2: Geographic coverage of surveys with randomly assigned man's questionnaire

The random assignment of the man's questionnaire to households is stratified by enumeration area. The treatment probability varies between 1/12 and 2/3 across

[19]The continuous DHS in Senegal and Peru are exceptions to this. They are carried out annually.

[20]We identify additional surveys with a man's questionnaire that is randomly assigned across households. We do not include these here because either their design differs in important ways from the one described in Section 3.2 or the available microdata does not lend itself to our analysis. Details are provided in Appendix A.2.1. We also exclude surveys that do not have national coverage.

surveys, but it is most frequently 1/2 (in 55% of surveys) or 1/3 (in 34% of surveys). The median duration of the man's questionnaire varies between 6 and 50 minutes across surveys, with the average man's questionnaire lasting 25 minutes.

In a subset of surveys (76), men and/or women in treatment households who are eligible for the individual questionnaire as well as children under the age of 5 are also eligible for biomarker collection. This typically amounts to a combination of HIV testing among eligible adults, anaemia testing among eligible women and children, and malaria testing and anthropometry among children. Men's biomarkers are collected in 58 of these surveys. In all of these cases, we estimate the joint impact of the man's questionnaire and biomarker collection.

Microdata for the identified surveys is obtained from the DHS (ICF, 1982-2022) and MICS (UNICEF, 2000-2022) online microdata archives. All variables required for the analysis are harmonised across datasets, as detailed in Appendix A.2.4.

Results. We find that household assignment to the man's questionnaire leads to the recording of a significantly lower number of eligible men in most surveys. Figure 3 plots the point estimates and 95% confidence intervals of the β coefficient from equation (3) relative to the control mean, sorted by magnitude across surveys. We estimate a statistically significantly negative impact in 130 out of 181 surveys (72%). For the remaining 51 surveys, our point estimates are mostly negative, but insignificant (36 surveys). Only for a single survey, we estimate a statistically significant positive effect. The average reduction in eligible men amounts to 7.7%. In 25% of surveys the reduction exceeds 9.3%, peaking at 23%.[21]

Surveys with longer man's questionnaires display more missing men. For example, an increase in the length of the man's questionnaire by 69 questions, corresponding to the difference between the 25th and the 75th percentile in the distribution of questionnaire length in our sample, is associated with an increase in missing eligible men by 2.2 percentage points, as shown in Figure A5a.[22]

Data collectors can achieve the observed reductions in the number of eligible men in treatment households in at least three ways. First, they can manipulate the eligibility criteria of household members such that they do not qualify for the man's questionnaire. Second, they can omit eligible men entirely from the household roster.

[21]Findings are robust to weighting households in each survey by their respective sampling weights (see Figure A4a). Table A4 provides pooled regression results (columns 1 and 2).

[22]Effects are also larger when biomarkers are collected with the questionnaire (see Figure A4b).

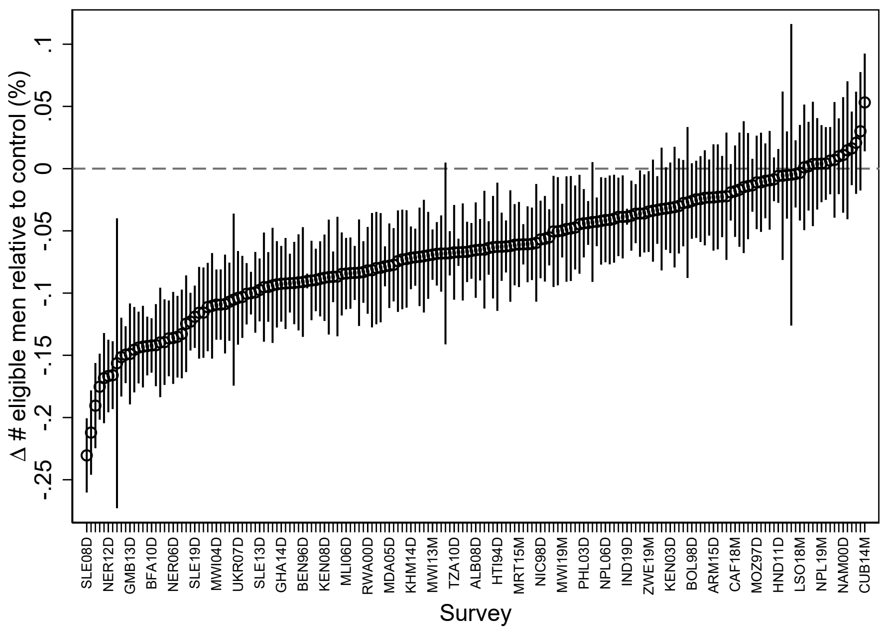


Figure 3: Effect of man's questionnaire on number of eligible men in the household

This figure displays estimates of β from equation (3) relative to the control mean where the outcome is the number of eligible men in the household. Circles indicate point estimates; bars indicate 95% confidence intervals. Surveys are sorted in ascending order of the point estimate. Survey labels are composed of country code, survey year and the first letter of the survey program (DHS/MICS).

Third, they can behave in ways that reduce the household response rate among eligible households with a large number of eligible men.

We find strong evidence in support of the first two margins, manipulation of eligibility criteria and omission of eligible men, but not the third. The number of ineligible men in treatment households exceeds the one in control households in many surveys, consistent with manipulation of eligibility criteria.[23] As shown in Figure 4, our point estimates are significantly positive for 56 surveys, significantly negative for 6 surveys and statistically insignificant in the remaining 119 surveys. Reassuringly, the total number of men in households is not positively affected by treatment in any surveys. Instead, it is either unaffected by treatment (94 surveys) or negatively impacted (87 surveys). The latter indicates that in many surveys, omission of eligible men from rosters is an important channel through which data collectors reduce their workload (see Figure A6a). Pooling all surveys in our sample, we estimate that age

[23]Ineligible men are those who do not qualify for the man's questionnaire because of their age, but are at least 10 years old. In the DHS, ineligible men additionally include all those of age 10 and above who did not sleep in the household in the night before.

displacement accounts for circa 30% of missing eligible men and omission for 70%.[24]

We find limited evidence of differential household response by household assignment to the man's questionnaire. First, note that household response rates in the surveys under study are very high. In fact, the average survey in our sample has a response rate of 97.4%.[25] Hence, there is limited scope for differential response. Second, MICS data allow us to explicitly test for balance in response.[26] We find that response is balanced between treatment and control in all but 5 out of 43 surveys (see Figure A6b). In all five cases, treatment is associated with marginally lower response rates, ranging between 0.3 and 1.4 pp. Hence, strategic manipulation of household responses does not appear to be an important margin of data collector response.

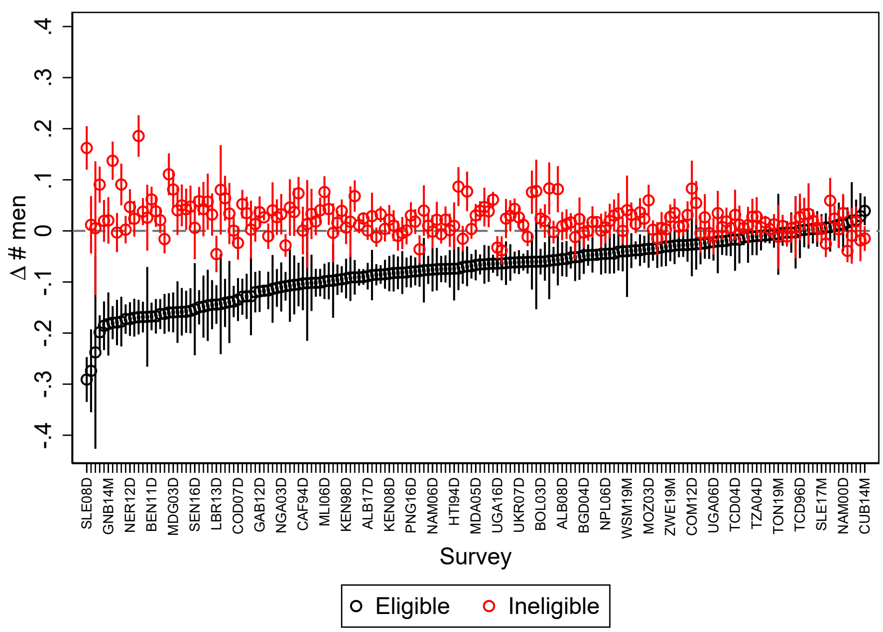


Figure 4: Effect of man's questionnaire on number of eligible and ineligible men

This figure displays estimates of β from equation (3) where the outcome is the number of eligible (black) and ineligible men in the household (red), respectively. Circles indicate point estimates; bars indicate 95% confidence intervals. Surveys are sorted in ascending order of the point estimate on the number of eligible men. Labels as in Figure 3.

It is worth noting that the administration of the man's questionnaire goes hand in hand with a change in data collectors. A strong emphasis on same-sex interviews in the DHS/MICS means that a male interviewer is required for households that are

[24]See Appendix A.3.2 for details on the decomposition into age displacement and omission.

[25]Response rates are sourced from the final reports of all surveys in our sample.

[26]The DHS program does not publish data on households that did not complete the interview. However, the MICS program provides this data for all but three surveys in our sample.

eligible for the man's questionnaire which is not the case for ineligible households. As a result, household rosters in treatment households are more likely to be administered by male data collectors. However, consistent with moral hazard, selection of data collectors into treatment cannot explain the reductions in the number of eligible men as estimates are barely affected by the inclusion of data collector fixed effects.[27]

4.2 Missing women

Empirical strategy. In the DHS and the MICS, women's responses to the woman's questionnaire are of central interest because they form the main focus of the two survey programs, namely the situation of women and children. Data collectors face substantially longer individual questionnaires with eligible women than with eligible men. In our sample of surveys, the median duration of the woman's questionnaire exceeds the median duration of the man's questionnaire in every single survey. On average, the woman's questionnaire is 16 minutes (64%) longer than the man's questionnaire. In conjunction with the results presented in the previous section, this raises serious concerns about endogenous sample selection of eligible women.

To assess the amount of missing women of eligible age, we cannot rely on the same identification strategy as for men because the woman's questionnaire is usually administered in all households, not just a random subset. We identify three (partial) exceptions to this rule, however, and leverage the random assignment in these three surveys to test if our results for men also hold among women.[28]

We complement these three cases with a systematic comparison of the number of female household members of eligible and ineligible age in DHS/MICS and contemporaneous population censuses. This is motivated by the fact that in the DHS and the MICS the number of questions to be administered to women of eligible age (typically aged between 15 and 49) is much larger than the number of questions to be administered to women outside this age range, but no such difference in question load between women of eligible and ineligible age exists in population censuses. This means that data collectors have a strong incentive to omit (or manipulate ages of)

[27] See Appendix A.3.3 for the effect of the man's questionnaire on data collector characteristics and within-data collector effects, and Appendix A.3.4 for effects on respondent characteristics.

[28] The three surveys are Ghana 2008 DHS, where the woman's questionnaire was only administered in a random subset of households; Namibia 2013 DHS and Gabon 2019 DHS, where a short version of the woman's questionnaire was administered to women aged 50 to 64 in a random subset of households (in addition a standard woman's questionnaire for women aged 15-49 in all households).

women of eligible age in the DHS and the MICS, but they have no such incentive in censuses. Hence, we can compare the average number of women of eligible and ineligible age in the household in the DHS/MICS and the census to test if survey samples contain fewer women of eligible age and (weakly) more of ineligible age.

Data. We form survey-census pairs by matching all DHS and MICS with population censuses conducted within two years of the survey. Since the MICS only records *de jure* household members, we ensure that censuses matched with MICS record all *de jure* members.[29] When comparing a DHS to a contemporaneous population census, we restrict the data to *de facto* household members because DHS statistics generated from individual questionnaires are based on *de facto* members only.[30] For 75 of the resulting census-survey pairs, we obtain microdata from IPUMS-International (Ruggles et al., 2024) or directly from national statistical offices.[31] See Table A3 for a complete list of the pairs and data sources, which cover 38 countries across Africa, Asia and Latin America, as shown in Figure A8.[32]

We confirm that the relative question load of eligible to ineligible women is close to one in all censuses, but much larger in the matched DHS and MICS. As shown in Figure A9, the relative question load varies between 1.0 and 1.5 across the matched censuses while it varies between 1.1 to 29.4 across the matched surveys.

Results. Exploiting the random assignment of the woman's questionnaire to households in three DHS, we find a sizable effect of the woman's questionnaire on the presence of eligible women in households in 2 out of 3 surveys - in line with our results for men presented in the previous section. Moreover, the effects of the woman's and the man's questionnaire are of the same order of magnitude within the same survey, as shown in Figure A10.

Comparing the number of eligible and ineligible women in the household in the DHS/MICS and contemporaneous population censuses paints a similar picture. Fig-

[29]For a subset of MICS-census comparisons, we restrict the data to *de jure* members that slept in the household last night because matching censuses do not include *de jure* members who are absent.

[30]For a subset of DHS-census comparisons, we restrict to *de facto* members that are usual members of the household because matching censuses do not include *de facto* members who are visitors.

[31]The authors wish to acknowledge all the statistical offices that provided the underlying data making this research possible. See Table A3 for a complete list of these.

[32]We exclude seven DHS-census pairs where eligibility for the DHS woman's questionnaire is conditional on having ever been married. To ensure comparability between census and survey data, we exclude collective dwellings from census data.

ure 5 illustrates that households in the DHS/MICS almost always contain fewer women of eligible age and more of ineligible age. In some cases, however, they contain more or less of both eligible and ineligible women, which points to level differences in the number of recorded household members that may arise from differences in the implementation of household rosters between the DHS/MICS and the census. Importantly, however, the difference in ineligible women between census and DHS/MICS is always at least weakly greater than the difference in eligible women. Thus, in relative terms, the DHS/MICS are under-recording eligible women throughout.

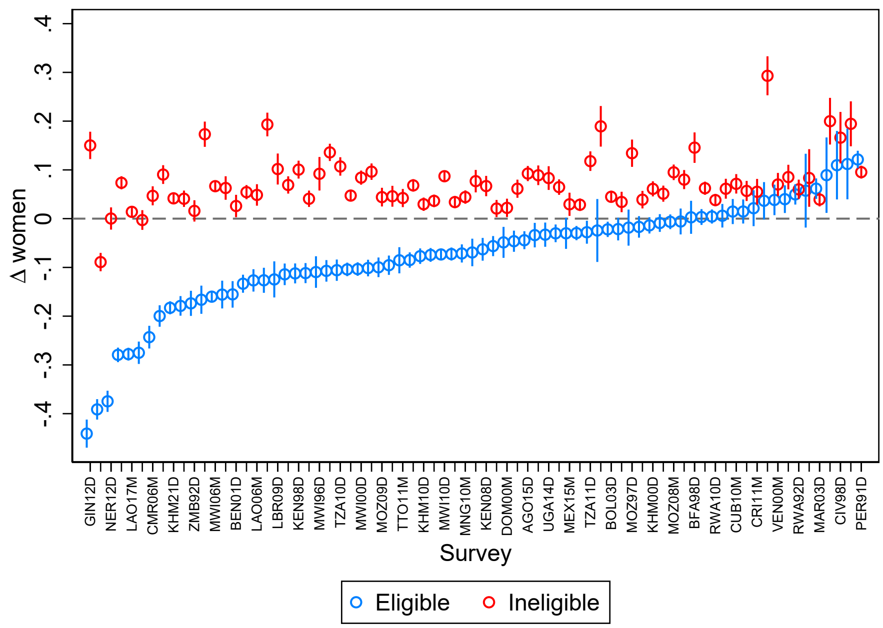


Figure 5: Missing and excess women in DHS/MICS relative to census

This figure displays estimates of β_3 from equation (6) where the outcome is the number of women of eligible (blue) and ineligible age (red) in the household. Circles indicate point estimates; bars indicate 95% confidence intervals. Surveys are sorted in ascending order of the point estimate on the number of eligible women. Labels as in Figure 3.

We provide bounds on the number of missing women that account for potential level differences in the recording of household members between censuses and DHS/MICS. We consider two extreme cases. First, to derive a lower bound, assume data collectors do not omit any eligible women from household rosters in the survey, but only engage in manipulation of age to displace eligible women across the eligibility thresholds. In this case, the number of missing eligible women in the survey relative to the census should equal the number of excess ineligible women, and any deviations from this equality would reflect level differences in the recording of household members between survey and census. Hence, the number of missing women is given by half of the difference-in-differences between the number of eligible and ineligible women in survey and census. Second, to derive an upper bound, assume data collectors do not engage in age displacement, but only omit eligible women from household rosters entirely. In this case, any deviation of the difference in the number of ineligible women between survey and census from zero reflects level differences in the recording of household members between survey and census. Thus, the number of missing eligible women is given by the entire difference-in-differences between the eligible and ineligible women in survey and census.[33]

Figure A11a displays the resulting bounds for missing women. We estimate a statistically significantly negative lower bound in 68 out of 75 surveys, ranging from 2% to 17%. In 11 surveys the lower bound exceeds 10%. The estimated upper bound is substantially larger (in absolute terms) and surpasses 10% in 45 surveys. This suggests that a substantial number of eligible women is screened out by DHS/MICS data collectors and never administered the woman's questionnaire. As with men, more women are missing in surveys with longer questionnaires (see Figure A5b).

To assess the bounds we construct for women, we turn to a subsample of DHS/MICS for which we have both a randomised man's questionnaire and a matched population census. This allows us to compare bounds of missing men (based on a survey-census comparison for households with a man's questionnaire) with our experimental estimates of the effect of the man's questionnaire. We find that the two approaches yield remarkably similar results (see Figure A11b). In 27 out of 33 surveys, the 99% confidence interval of the experimental estimate overlaps with the range of estimates delimited by the bounds. In the remaining cases, the experimental estimate falls short of the lower bound, mostly due to timing differences between survey and census.

To facilitate the comparison of our estimates across surveys and sex, we normalize our estimates of missing men and women relative to the length of the individual questionnaires faced by eligible household members. To this end, we compute the elasticity of the number of recorded eligible household members with respect to question load. We define this elasticity as the relative reduction in the number of eligible household members over the relative increase in question load for eligible household members. We measure the question load by the total number of questions listed in

[33]See Appendix A.4.1 for details on the estimation of the bounds.

the questionnaires that a household member is eligible for.[34]

We find similar elasticities for men and women. The elasticity for men estimated from the random assignment of the man's questionnaire is on average -0.010, with variation across surveys between -0.001 at the 10th percentile and -0.021 at the 90th percentile. The elasticity for women estimated from the survey-census comparison is on average -0.008, ranging between -0.002 at the 10th percentile and -0.027 at the 90th percentile. In 33 surveys where we can estimate both of these elasticities, they are quantitatively similar and rarely significantly different from each other.[35]

4.3 Mechanisms

Effort cost. The theoretical framework introduced in Section 2 suggests that endogenous sample selection is more severe if the effort cost of recording household members is larger. We test this prediction using variation in temperature and humidity as a source of exogenous variation in effort cost. Motivated by experimental evidence of their negative impact on human physiology and performance (Pilcher et al., 2002; Seppanen et al., 2005; Cui et al., 2013), we hypothesise that reductions in eligible men will be more pronounced at low and high temperatures relative to intermediate temperatures.

We use wet bulb temperature as our preferred measure of temperature, following the recent literature in economics (Geruso & Spears, 2018; LoPalo, 2023). Wet bulb temperature accounts for relative humidity which interacts with temperature in the generation of heat stress. Note that wet bulb temperature is lower than dry bulb temperature unless relative humidity is 100% – in which case the two are equal.[36]

We estimate how the effect of household assignment to the man's questionnaire varies with temperature changes within survey cluster and data collector, adopting a semi-parametric specification to allow for non-linearities in the effect of temperature:

[34] Formally, the elasticity is $\varepsilon = (\beta/m_c)/[(q_{eligible}/q_{ineligible})-1]$ where β is the effect of household assignment to the man's questionnaire on the number of eligible men as estimated in equation (3), m_c is mean number of eligible men in control households and q_i is the number of questions that an (in)eligible member has to be asked. Appendix A.2.3 provides details on how questions are counted.

[35] See Figure A12a for survey-level elasticity estimates for men and women, and Figure A13 for the subset of surveys where we can estimate both in the same survey.

[36] Moderate conditions, such as a dry bulb temperature of 22°C and a humidity of 45%, correspond to wet bulb temperatures of around 15°C. Wet bulb temperatures of 25°C capture extreme heat, corresponding to dry bulb temperatures of 35°C at 40% humidity or 30°C at 65% humidity.

$$y_{icdrt} = \sum_j \beta_j (T^j_{ct} \times MQ_{ic}) + \sum_j \gamma_j T^j_{ct} + \eta Precip_{ct} + \mu_c + \theta_d + \lambda_r + \epsilon_{icdrt} \quad (4)$$

where y_{icdrt} is the number of (in)eligible men in household i interviewed on the r-th day of data collection in survey cluster c by data collector d on date t. MQ_{ic} is an indicator variable that takes value one if household i in cluster c is eligible for the man's questionnaire, and zero otherwise. T^j_{ct} is an indicator that takes the value one if the daily average wet bulb temperature in cluster c on date t falls into temperature bin j. The coefficients of interest are β_j capturing the reduction in the number of eligible men due to household assignment to the man's questionnaire in the different temperature bins. We control for precipitation $Precip_{ct}$, survey cluster fixed effects μ_c, data collector fixed effects θ_d and fixed effects for the day of data collection in the survey cluster λ_r. Standard errors are clustered at the survey cluster level.

We construct the daily average wet bulb temperature in each survey cluster on each survey date following LoPalo (2023). We source global daily weather information for each 0.25 degree latitude/longitude increment from the Princeton Meteorological Forcing Dataset.[37] For each survey cluster, we set the wet bulb temperature to the average across the four surrounding grid points, weighting by the inverse distance between the cluster and each grid point. This allows us to link 17,716 survey clusters from 49 of the surveys in our sample to weather information.[38]

In line with our hypothesis, we find an inverse U-shaped relationship between the treatment effect on eligible men and wet bulb temperature, as shown in black in Figure 6. Treatment effects are smallest at wet bulb temperatures between 10°C and 20°C where 0.08-0.09 eligible men are missing. We find substantially more missing men at higher wet bulb temperatures: approximately 0.12 eligible men are missing between 20°C and 25°C, and 0.15 above 25°C. We also observe somewhat more missing men at wet bulb temperatures below 10°C, although estimates are noisier. Consistent with the idea that missing eligible men are partially driven by age displacement, results for ineligible men (shown in red) mirror those for eligible men, with U-shaped estimates in temperature, but smaller in absolute values than for eligible men, as

[37]The original data is 3-hourly. We work with the average wet bulb temperature across the 8 daily readings. For more details on the weather dataset see LoPalo (2023).

[38]Clusters from the remaining surveys cannot be matched because, first, the Princeton Meteorological Forcing Dataset only covers the time period until 2010; second, GPS coordinates of survey clusters are not available. We provide robustness using alternative weather data in Appendix A.5.1.

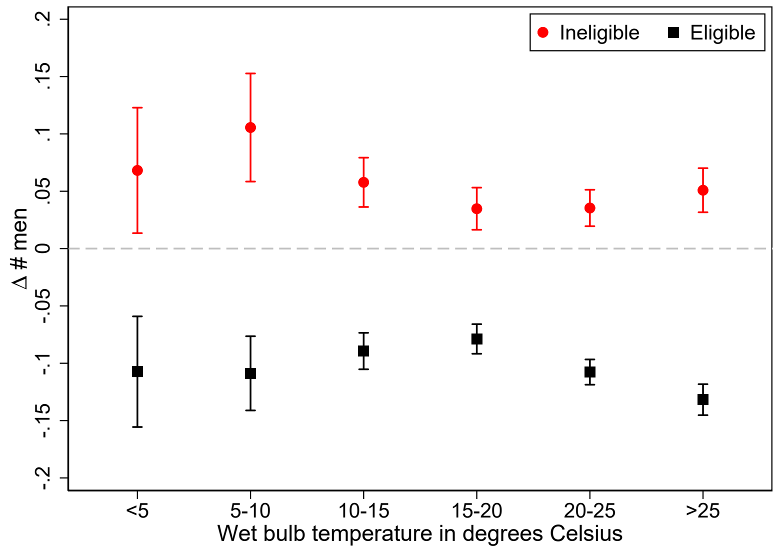


Figure 6: Missing men by wet bulb temperature bin

This figure displays the estimates of the regression coefficients β_j from specification (4) across temperature bins. Bars indicate 95% confidence intervals. Standard errors are clustered at the survey cluster level.

expected if both age manipulation and omission are present.

Interestingly, the inverse U-shape we estimate is similar to the relationship between ambient temperature and human productivity estimated by others (Cai et al., 2018; LoPalo, 2023), suggesting that data collectors partially compensate for productivity losses experienced at high temperatures by reducing the number of eligible household members. More broadly, our findings indicate that higher temperatures may aggravate moral hazard problems in the workplace.

Detection probability. Our theoretical framework further predicts that screening out of eligible men will be less severe if the probability of getting caught doing so is higher. To examine this prediction, we study the correlation between quality control measures and the share of missing eligible men across surveys. These measures include systematic backchecks, also referred to as mandatory re-interviewing or audits, data monitoring through field check tables and the use of tablets which can facilitate quality control. Consistent with our framework, we find that the share of missing men is lower in surveys that conduct systematic backchecks: the average share of missing eligible men is 2.6% in surveys with backchecks compared to 8% in those without. In contrast, field check tables and tablets are only weakly associated with the number of eligible men – potentially because they do not provide a direct means to detect roster manipulation.[39]

We provide further evidence in line with our theoretical framework by showing that the effect of the man's questionnaire on missing men is larger for age groups for which age manipulation is plausibly harder to detect for supervisors and easier to negotiate with survey respondents. First, we demonstrate that effects are larger among age groups close to the lower and the upper age eligibility threshold, for whom age manipulation is arguably easier because it requires a smaller deviation from the truth. We estimate that 9.3% and 7.9% of eligible men within 10 years of the lower and upper age threshold, respectively, are missing as a result of the man's questionnaire. In contrast, for the age group in-between, the share of missing men is only 4.7%.

Second, we compare effect sizes for men of round and non-round ages. This is motivated by a large literature that has shown significant age heaping at round ages (ending in 0 or 5) in many low- and middle-income countries likely because respondents to household rosters do not know the exact ages of all household members (Myers, 1940; Bachi, 1951).[40] We hypothesise that manipulating the age of household members whose exact age is not known by respondents is easier. Consistent with this, we find that the effect of the man's questionnaire is much larger among men of round ages (13.7%) than non-round ages (4.4%) – see Figure A15. An alternative interpretation of this finding is that household members whose age the respondent is not sure about are easier to omit because their household affiliation is ambiguous.

Other testable implications. Our framework yields further testable implications that are supported by the data. First, the negative effect of the man's questionnaire on the number of eligible men becomes more pronounced as surveys progress, consistent with a decreasing incentive to record household members as the continuation value of data collector contracts falls. Second, endogenous sample selection becomes more pronounced as data collection progresses within survey clusters, consistent with the importance placed on adherence to pre-determined travel schedules of data collection teams between clusters, as described in Section 3. Third, the share of missing men is larger in large households, as proxied for by the existence of polygamy in the household, again consistent with a notion of time pressure as an important driver of

[39]See Table A5 for the corresponding regression results.

[40]This is well-documented for the DHS (Pullum, 2006) and evident in Figure 1.

roster manipulation.[41]

5 Selection

Eligible men and women imply a high effort cost for data collectors because they have to be administered lengthy individual questionnaires. In the previous section, we established that in many DHS and MICS a significant number of eligible men and women are excluded from individual questionnaires as a result of manipulation by data collectors. Consequently, these men and women are missing from the data collected in individual questionnaires, and all statistics derived from that data. These statistics, however, are the core output of the DHS and the MICS programs. They include crucial information on topics such as fertility, maternal health, HIV, marriage and domestic violence. Thus, our findings raise an important question: does the remaining selected sample of individuals differ systematically from the full sample that would have been observed in the absence of manipulation by data collectors?

We compare average characteristics of eligible men listed in households with and without a man's questionnaire to assess selection along observable dimensions. To this end, we regress household-level assignment to the man's questionnaire on individual-level characteristics recorded in the household roster, which are observable for all men independent of their household's assignment to the man's questionnaire.

We find that manipulated samples differ systematically from complete samples. Figure 7 reports estimates from six separate pooled regressions of assignment to the man's questionnaire on different individual characteristics recorded in household rosters: eligible men recorded in households assigned to the man's questionnaire are on average 3% less to be likely distantly related to the head of their household (relative to eligible men in control households), 3% less likely to never have attended school, 3.5% less likely to be chronically ill, 3.5% less likely to never have married and 12.7% less likely to be poor, but not differentially likely to be disabled. In turn, this implies that the screened out eligible men are more distantly related to their household heads, less educated, less healthy, less likely to have ever married and poorer.[42]

Selection patterns for women are very similar to those of men. All selection results for women are shown in Appendix A.6, including a graphical comparison to men.

[41] See Appendix A.5.3 for evidence on all findings on other testable implications.

[42] We show in Appendix A.6 that selection results are robust to jointly including individual characteristics, and provide survey-by-survey estimates for the most commonly observed characteristics.

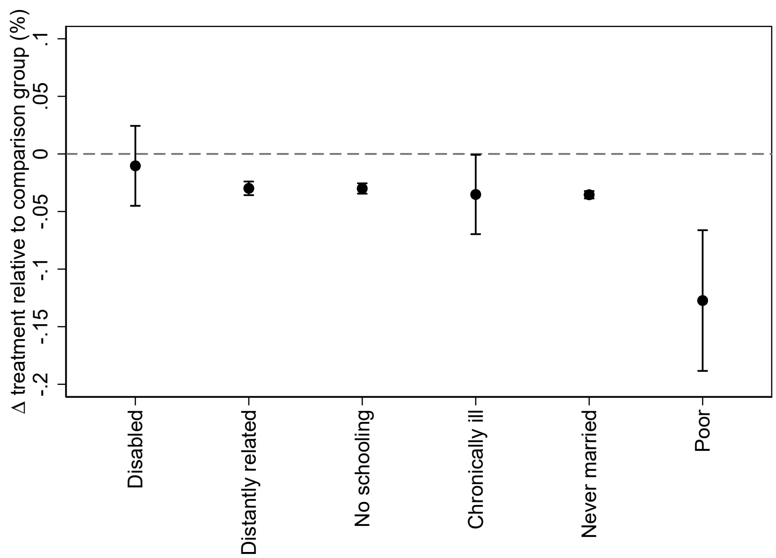


Figure 7: Selection on observables

This figure displays the estimated coefficients from separate pooled regressions of household assignment to the man's questionnaire on individual characteristics. Variable construction is described in Appendix A.2.4 and regression results reported in Appendix A.6. Bars indicate 95% confidence intervals. Standard errors clustered at the household level.

One interpretation of the observed selection pattern is that data collectors predominantly screen out eligible individuals that are at the margins of their respective households. In other words, data collectors screen out precisely those household members for which they have discretion because household definitions are sufficiently ambiguous, with rosters typically instructing data collectors to list all 'usual members' (plus visitors that slept in the household last night in the case of the DHS). Moreover, omitting (or manipulating ages of) marginalised members of a household is plausibly less likely to face opposition from respondents such as the household head. In terms of our theoretical framework in Section 2, the probability of detection is likely lower.

Reassuringly, we find few signs of selection in surveys with low missingness (see Figure A18). Eligible men in treatment and control households look similar on observables in surveys with few missing men, but differ in surveys with more missing.[43]

[43] Analogous evidence for missing women is shown in the right hand side panels of Figure A18.

6 Implications for policy and research

6.1 Implications for policy

The documented selection on observables implies that endogenous sample selection does not only lead to a decline in precision of estimates as a result of sample size reductions. It also leads to bias in aggregate statistics. How important is this bias? In this section, we address this question focusing on two key survey outputs – fertility estimates and Sustainable Development Goal (SDG) indicators.

Fertility estimates. The DHS and MICS are a key data source on fertility in low- and middle-income countries. The large number of top demography papers citing the DHS and the MICS is evidence of this. Between 2013 and 2017, 15.4% of all papers published in the two top journals *Demography* and the *Population and Development Review* cited the DHS or the MICS.[44] Work on fertility in the field of economics also heavily relies on the two household survey programs (Vogl, 2016; Chatterjee & Vogl, 2018; Rossi, 2018; Dupas et al., 2023; Zipfel, forthcoming). Additionally, fertility data from the two programs is a key input for national health, family planning and education programs, not least due to the weakness of vital registration systems in large parts of the world. In fact, the DHS and the MICS are considered the only reliable data source on fertility in many contexts.

We pursue two complementary approaches to gauge the effect of endogenous sample selection on fertility estimates from the DHS. First, we leverage the random assignment of the man's questionnaire. In the absence of information on the fertility of men in households without a man's questionnaire, we show that in the pooled data, the number of biological children that eligible men live with in their household, is 4.8% larger in treatment than control households (black *Full sample* in Figure 8).[45]

Second, we compare the total number of live births reported by women aged 15 to 49 in the DHS and contemporaneous population censuses, which we consistently

[44]Figures based on keyword search for "Demographic and Health Survey(s)" and "Multiple Indicator Cluster Survey(s)" across all fields on JSTOR.

[45]Since the fertility of men is only elicited in the man's questionnaire, we do not observe fertility of men in control households. We overcome this limitation by computing the number of biological children each eligible man lives with as a proxy of fertility in both treatment and control households, using the parent survival module in the household roster (available 126 out of 135 DHS in our sample). To obtain nationally representative figures, we apply sampling weights.

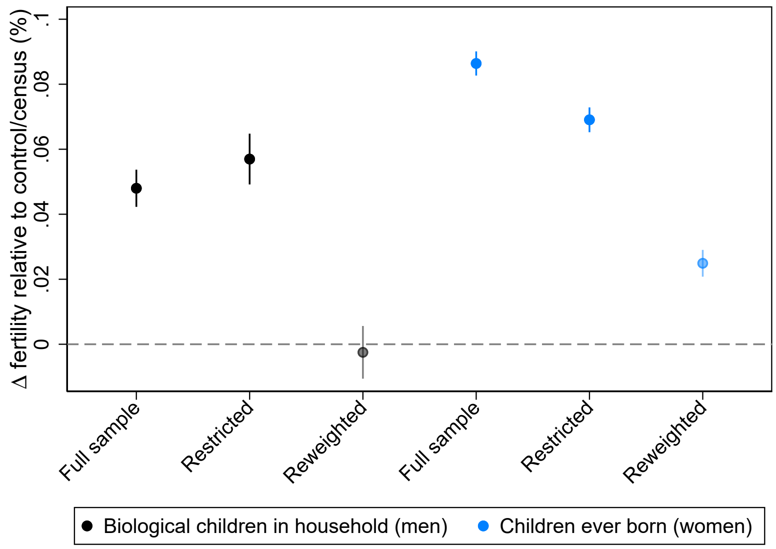


Figure 8: Bias in aggregate fertility statistics and selection correction

This figure displays the estimated coefficients from separate pooled regressions of fertility measures on assignment to the man's questionnaire. Bars indicate 95% confidence intervals. Standard errors clustered at the household level.

observe in 50 DHS-census pairs.[46] We find that this metric is on average 8.6% larger in DHS than in censuses (blue *Full sample* in Figure 8). If we focus on completed fertility instead and restrict the sample to women aged 45-49, this estimate is even larger, amounting to 9.5% (not shown).

A caveat to the DHS-census comparison is the fact that information on live births is not collected in the same way across the two instruments. While the DHS typically asks women for their entire birth history, censuses only ask for the total number of children ever born. It is thus conceivable that the more thorough approach of the DHS is (partially) accountable for the difference in recorded live births. To address this concern, we reweight women in the DHS on four observable characteristics (age, years of schooling, relationship to the household head and marital status) using propensity score reweighting and re-run our DHS-census comparison. Because the full set of variables for reweighting is not available for all DHS-census pairs, this analysis is restricted to 37 pairs. As Figure 8 (blue) shows, in this subsample live births recorded in the DHS exceed those in the census by 6.9%. After reweighting, the gap shrinks to 2.5%, indicating that selection on observables *alone* can account for a large part of the observed difference in live births. The remaining difference can be due to many factors, including selection on unobservables, differences in the collection of the

[46]See Appendix A.2.4 for details on the survey-census harmonisation of this information.

outcome variable, differences in the collection of the observables used for reweighting and the time gap between survey and census. Interestingly, we find that the same reweighting procedure entirely removes the difference in co-residing biological children between men in treatment and control households (black *Reweighted* in Figure 8), suggesting that differences in the collection of relevant variables are more important than selection on unobservables in our context.[47]

In conclusion, our findings indicate that data collectors are more likely to screen out eligible men and women with less children, in line with the relative absence of the never married from treated households documented in Section 5. Ultimately, this leads to an upward bias in DHS fertility measures of approximately 5%. However, reweighting on observables goes a long way in correcting this bias.

Monitoring progress towards the SDGs. The DHS and the MICS are a key data source for a wide range of SDG indicators. Motivated by the results from the previous section, we explore how endogenous sample selection distorts reported indicators by reweighting the underlying samples of eligible women such that they closely resemble women in contemporaneous population census data along observable dimensions, following the approach explained in the previous section. In addition to propensity score reweighting, we employ raking as an alternative reweighting approach.[48]

We focus on six indicators derived from the woman's questionnaire. As shown in Figure 9, reweighted statistics differ from reported ones, but the magnitude of these differences varies greatly across SDG indicators. Concentrating on the results from propensity score reweighting, we find that at the median (75th percentile) the reweighted share of births attended by skilled personnel is 1% (3%) lower than reported, reweighted under-5 and neonatal are 1% (2%) higher, child vaccination is 1% (2%) lower and child marriage 7% (12%) lower. There is no meaningful difference between the reported and the reweighted share of women making their own decisions on contraceptive use, and results from raking are overall very similar across to those from propensity score reweighting.

We view this evidence as suggestive of the direction and magnitude of distortion in SDG indicators caused by endogenous sample selection. However, it is important to note that distortions may in fact be larger as we are only correcting for selection

[47]See Appendix A.7 for details on the propensity score reweighting approach.

[48]See Appendix A.7 for details on raking. The SDG indicators that can be estimated using the DHS are listed under https://dhsprogram.com/Topics/upload/SDGs_in_DHS_15Mar2023.pdf.

on four observables here.

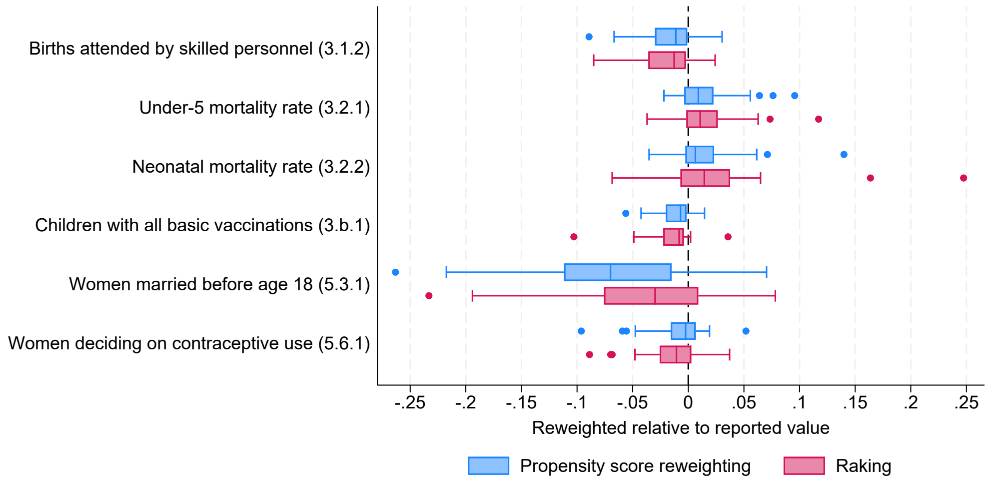


Figure 9: Reweighted Sustainable Development Goal (SDG) indicators

This figure displays the distribution of the difference between reweighted and reported statistics relative to reported ones across surveys for each SDG indicator and reweighting approach in the form of a box-whisker plot. One outlier with large differences, TZA AIS 2011, is not shown.

6.2 Implications for research

How does endogenous sample selection affect economic analysis and research? Endogenous sample selection poses a threat to the causal identification of treatment effects if it is correlated with treatment. To assess the relevance of this concern, we study the effect of prominent causes of economic and social development on endogenous sample selection. We focus on three broad sets of causes – climate shocks, institutions and economic origins – and leverage empirical strategies from existing studies for identification. Geographically, the center of our analysis is Sub-Saharan Africa where the DHS/MICS have been a key input to a lot of influential research.

Climate shocks. A rapidly growing literature uses DHS data to examine the effect of climate shocks on a wide range of outcomes, including health (Burke et al., 2015; Geruso & Spears, 2018; Thiede & Strube, 2020; Nagata et al., 2021; Blom et al., 2022; Le & Nguyen, 2022; Dimitrova et al., 2024; Gray & Thiede, 2024; He et al., 2025), fertility (Norling, 2022; Abdel Ghany et al., 2026), marriage (Corno et al., 2020), domestic violence (Epstein et al., 2020), consumption (Dimitrova, 2021) and wealth (Thiede, 2014). In this literature, climate shocks are typically considered exogenous events and associations between shocks and outcomes are interpreted causally. However, if outcome data is collected by humans on the ground, selection into the sample can itself be affected by climate shocks insofar as these shocks impact the behaviour of data collectors. This, in turn, undermines the causal interpretation of the association between shock and outcome even if the shock is truly exogenous.

In this section, we examine how drought and heat in Sub-Saharan Africa affect the extent to which data collectors screen out eligible men in treated households. To this end, we construct a grid-cell-year panel of weather data and link it to the DHS in our sample using survey cluster coordinates, as detailed in Appendix A.8.1. Following the literature, we define a drought as calendar year rainfall below the 15th percentile of a grid cell's long-run rainfall distribution. Heat is measured by the number of days with a maximum temperature above 30°C in the calendar year of the survey following Somanathan et al. (2021). To identify the causal effects of these shocks, we adopt an empirical approach akin to Burke et al. (2015) and Corno et al. (2020), and estimate how household assignment to the man's questionnaire interacts with changes in rainfall and temperature within grid-cell using the following regression model:

$$y_{igct} = \beta_0 + \beta_1 MQ_{igct} + \beta_2 Shock_{gt} + \beta_3 (MQ_{igct} \times Shock_{gt}) + \mu_{cg} + \tau_{ct} + \epsilon_{igct} \quad (5)$$

where y_{igct} is the number of eligible men in household i in grid cell g in country c in year t, MQ indicates household assignment to the man's questionnaire and $Shock$ is either an indicator for drought or the above measure of heat.

We find that climate shocks affect sample selection. In line with the positive impact of wetbulb temperature, a combined measure of temperature and humidity, on missing eligible men documented above, we find that heat aggravates endogenous sample selection and drought alleviates it. As shown in Table 1, a one-standard-deviation increase in heat leads to an increase in missing eligible men due to treatment from 0.06 to 0.082. Droughts, on the other hand, lead to a reduction in missing eligible men from 0.094 to 0.06.

Institutions and historical legacies. A large literature demonstrates the persistent effects of historical institutions on development today. We test whether such institutions also affect the extent of endogenous sample selection in present-day data collection, focusing on four important examples from the literature: colonial concessions (Lowes & Montero, 2021a), colonial medical campaigns (Lowes & Montero, 2021b), slave trade (Nunn, 2008; Nunn & Wantchekon, 2011) and segmentary lineage (Moscona et al., 2020). Following the empirical approaches of the original studies, we exploit spatial variation in the exposure to these institutions to estimate their effect on the extent of missing eligible men due to the man's questionnaire.[49]

We find that the former three exploitative colonial institutions all increase the number of eligible men missing in households with a man's questionnaire. This suggests that these institutions had a lasting negative effect on the trust of citizens in the state, creating a situation in which data collectors interested in screening out as many eligible men as possible encounter respondents who have little desire to be recorded. As Table 1 shows, the effect of colonial rubber concessions in the DRC, where men would be rounded up and enlisted for forced labor, is particularly large, increasing the number of missing men from 0.12 to 0.28. The effects of colonial medical campaigns in West Africa forcing millions to receive injections of medications with serious side effects as well as the effects of slave exports are somewhat smaller but still significant. We estimate that a one-standard deviation increase in these treatments is associated with an increase in missing men by 0.2 and 0.1, respectively.

Finally, we find that the organisation of ethnic groups around segmentary lineages is associated with less missing eligible men across adjacent ethnic homelands in Sub-Saharan Africa, although the coefficient of interest is not statistically significant at conventional levels (p-value of 0.105).

Economic origins. A small literature traces the roots of social norms today back to economic origins. Following two well-known contributions to this literature, we investigate whether traditional plough agriculture (Alesina et al., 2013) and traditional pastoralism (Becker, 2024) are not only predictive of present-day gender norms, but also of endogenous sample selection in the DHS. We find that plough agriculture is associated with more missing eligible men across ethnic homelands in Sub-Saharan Africa while pastoralism is associated with less, although not statistically significantly so (p-value of 0.14).[50]

Summary. The reported effects of climate shocks, institutions and economic origins on sample selection caution against the causal interpretation of correlations between these treatments and outcome data collected by humans in the field. If treatment is correlated with sample selection, treatment effects are not point-identified. However,

[49] See Appendix A.8 for details on the separate empirical approaches.

[50] Additional details are provided in Appendix A.8.

Table 1: Causes of endogenous sample selection

Theme (1)	Treatment (2)	Design (3)	MQ (4)	MQ x Treatment (5)	No. hhs (6)	No. svys (7)	Related literature (8)
Climate shocks	Heat *(continuous)*	TWFE	-0.06 [0.000]	-0.022 [0.000]	847646	76	Geruso et al. (2018), Somanathan et al. (2021)
	Drought *(binary)*	TWFE	-0.094 [0.000]	0.034 [0.000]	898776	76	Burke et al. (2015), Corno et al. (2020)
Institutions	Colonial concessions *(binary)*	RDD	-0.118 [0.030]	-0.161 [0.063]	2803	2	Lowes et al. (2021a)
	Medical campaigns *(continuous)*	OLS	0.001 [0.948]	-0.017 [0.001]	82366	9	Lowes et al. (2021b)
	Slave exports *(continuous)*	OLS	-0.082 [0.000]	-0.01 [0.021]	503695	65	Nunn et al. (2011)
	Segmentary lineage *(binary)*	RDD	-0.09 [0.000]	0.036 [0.105]	122419	41	Moscona et al. (2020)
Economic origins	Plough agriculture *(binary)*	OLS	-0.092 [0.000]	-0.052 [0.000]	861278	76	Alesina et al. (2013)
	Pastoralism *(binary)*	OLS	-0.096 [0.000]	0.032 [0.140]	861278	76	Becker (2024)

This table summarises how treatments of interest interact with household assignment to the man's questionnaire to determine the number of missing eligible men. Columns (4) and (5) report estimates of coefficients β_1 and β_3 from specification 5. For binary treatments, see column (2)), we provide the interaction coefficient; for continuous treatments, we multiply the estimated interaction coefficient by the standard deviation of the treatment variable. P-values in square brackets. In column (8), we list closely related papers motivating the design indicated in column (3). "TWFE": two-way fixed effects, "RDD": regression discontinuity design, "OLS": Ordinary Least Squares. See Appendix A.8 for further details.

it may still be possible to derive meaningful bounds following Lee (2009) and more recent work building on this classic paper (Gerard et al., 2020; Honoré & Hu, 2020).

7 External validity

The theoretical framework underpinning our analysis is generally applicable to human-collected data. In this section, we provide graphical evidence indicative of endogenous sample selection in a wide range of data products, confirming its broad relevance.

7.1 Household surveys beyond the DHS and the MICS

In many household surveys, household rosters serve a screening purpose with eligibility of household members for subsequent interviewing conditional on recorded demographics. Living standards, labour force, time use and education surveys all share this feature: eligibility for lengthy survey modules depends on the age of household members reported in the roster. In Figure A20, we present four examples to illustrate how this translates into endogenous sample selection: the Zambian Living Conditions Monitoring Survey (LCMS) 2015, the Tanzanian Integrated Labour Force Survey (ILFS) 2014, the Ghanaian Time Use Survey (TUS) 2009 and the Comprehensive Modular Survey on Education in the Indian National Sample Survey (NSS) 2007. In all cases, age distributions clearly display excess mass just outside eligible age windows and missing mass inside these windows, suggesting that endogenous sample selection affects a wide array of existing household surveys.

Endogenous sample selection is of particular concern in longitudinal surveys since initial selection into the sample affects the representativeness of all subsequent survey rounds. We report evidence consistent with endogenous sample selection from a widely used longitudinal survey in the US, the National Longitudinal Survey of Youth 1997 (NLSY97). Households screened for this survey in 1997 show pronounced under-coverage of youths of eligible ages (12-16) compared to households in the contemporaneous Current Population Survey (CPS) from March 1997 (Figure A21a). This example also highlights the external validity of our findings beyond low- and middle-income country contexts.[51]

7.2 Firm surveys

Endogenous sample selection is not confined to household surveys. Firm censuses and surveys may likewise generate incentives for data collectors to endogenously omit or displace. Establishment size, for example, is frequently used as an eligibility criterion for additional questions.

Below, we show how endogenous sample selection interferes with data collection and affects the resulting data of an important firm census: the Indian Economic Census (EC), the world's largest. The Indian EC is heavily used in economics, featuring in at least six top general interest publications over the last five years alone. Yet, its design creates an incentive for data collectors to manipulate firm size. The EC aims to record all formal and informal non-farm businesses in India. To this end, data collectors visit all buildings in the entire country, record the firms found therein and their basic characteristics, including the total number of employees. Subsequently,

[51] An alternative explanation for the under-coverage is households hiding eligible youths although it is unclear to what extent they were aware of the exact age thresholds for eligibility at the time of screening. See Appendix A.9.1 for further details.

additional information is collected for firms above a given size threshold.

Curiously, the eligibility threshold for additional data collection shifted between consecutive censuses and enables us to reveal firm size manipulation by data collectors. In fact, no size threshold existed in 1998 – the amount of information collected about firms was independent of their size. In contrast, the 2005 EC introduced a requirement to complete an address slip for all firms employing 10 or more workers (Ministry of Statistics and Programme Implementation, India, 2005). The subsequent 2013 EC adjusted the eligibility threshold downward to a firm size of 8, while slightly expanding the additional information collected for each firm above the threshold.[52]

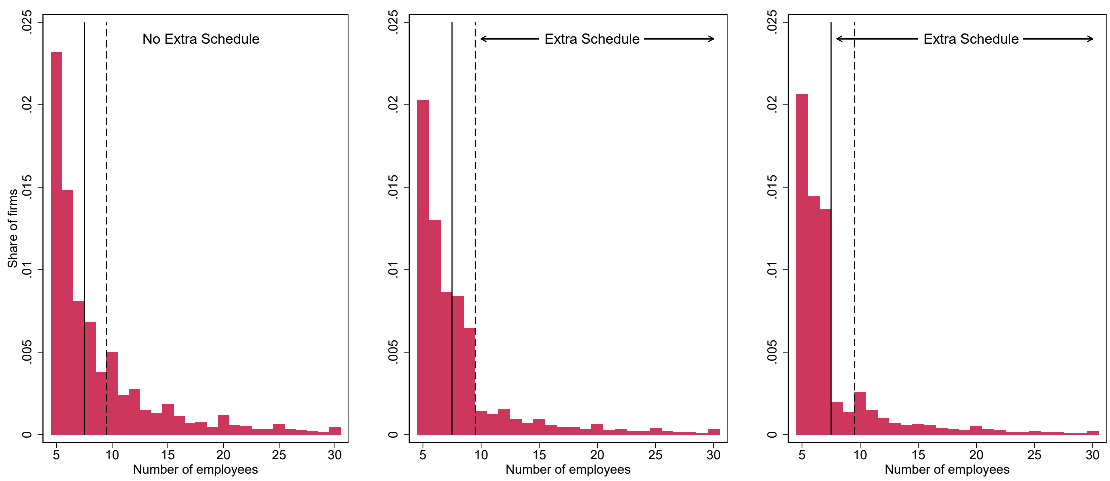


Figure 10: Firm size distribution in the Indian Economic Census 1998, 2005 and 2013

This figure plots the firm size distribution in the 1998, 2005 and 2013 Indian EC, on a log scale. Vertical lines indicate firm size thresholds above which data collectors had to complete additional schedules in 2005 (dashed line) and 2013 (solid line). Note that there was no variation in the number of schedules to be completed by firms size in 1998.

Figure 10 illustrates the resulting firm size distributions across years. It uncovers bunching of firms below the respective eligibility thresholds of 10 and 8 employees in 2005 and 2013, respectively, and no sign of bunching in the 1998 firm size distribution.[53] Moreover, bunching moves in accordance with the change in threshold between 2005 and 2013. We interpret this as evidence of firm size manipulation by data collectors.[54]

[52] We provide additional information on eligibility thresholds and question load in Appendix A.9.2.

[53] In results available from the authors, we confirm that the 1998 EC firm size distribution follows a power law, a regularity observed in many countries. In contrast, 2005 and 2013 firm size distributions violate this power law with excess (missing) mass to the left (right) of the respective size threshold.

[54] There are various regulations in India that apply to firms of size 10 and above which may contribute to bunching at firm size 9 in 2005 (Amirapu & Gechter, 2020). However, none of these

8 Discussion

8.1 Information-bias trade-off in data collection

The evidence presented in the preceding sections points to a fundamental trade-off in the collection of data between the amount of information collected from respondents on the one hand, and bias in the sample of respondents as a result of screening out by data collectors on the other hand. The elasticity of sample size with respect to question load of approximately -0.01 reported in Section 4.2 indicates that administering an individual questionnaire with as many individual questions as the household roster leads to the screening out of about 1% of respondents eligible for individual questionnaires. While this elasticity may seem low, it is important to recall that individual questionnaires in the DHS/MICS are typically multiple times longer than household rosters, thus leading to a substantial share of missing eligible individuals.

To put this cost of asking (for) more into context, we provide an illustration from the history of the DHS and MICS programs: since the 1990s, the number of questions asked in both survey programs proliferated with the average length of the man's questionnaire in our sample nearly doubling from 103 to 205 questions (as shown in Table A8, columns 1 and 2). At the same time, the elasticity of sample size with respect to question load, if anything, fell (columns 3 and 4). The combination of proliferation in question load and weakly falling elasticity, however, led to more missing men over time (columns 5 and 6), rising from 6.1% in the 1990s to 8.9% today. This upward trend in missing individuals is remarkable given simultaneous technological and managerial advances in data collection. It provides a cautionary tale that the ever-increasing demand for more information by researchers and policymakers comes at the cost of growing bias in sample composition due to endogenous selection.

8.2 Remedies

The widespread nature of endogenous sample selection and its implications for research and policy demand remedies. We provide an overview of potential countermeasures below, and call for further research to explore cost-effective alternatives.

Detecting endogenous sample selection. The ability to detect endogenous sample selection and measure its extent is critical to any attempt to effectively address

can explain the drop in firms of sizes 8 and 9 between 2005 and 2013.

it. Depending on data availability, different approaches can be pursued to this end. First, it is advisable to check for bunching of survey respondents at discontinuities in question load, either graphically or using formal tests for discontinuities in density. Second, measurement can be achieved by comparing distributions of demographic characteristics that determine survey eligibility with population distributions where these are available, as demonstrated for women in the DHS/MICS in Section 4.2. Third, random assignment of question load allows for precise quantification of the extent of endogenous sample selection, as shown in Section 4.1.

Correcting for endogenous sample selection. Since endogenous sample selection is an inherent feature in many existing datasets, it is essential to understand how it can be corrected for *ex post*. One potential avenue is to re-weight existing data suffering from endogenous selection along observable dimensions such that it more closely resembles population parameters estimated from undistorted data, as showcased in Section 6.1. We leave the exploration of other approaches, such as the use of selection models, for future research.

Preventing endogenous sample selection. How could endogenous sample selection be prevented *ex ante* in future data collection? As shown in Section 4.3, systematic backchecks are associated with less missing eligible men in the DHS/MICS, suggesting that this form of quality control limits endogenous sample selection. Other approaches beyond improved monitoring proposed by practitioners include a reduction in question load differentials across household members or firms, and the separation of the screening stage of data collection from the interview stage, by assigning the responsibility for each stage to a different data collector. Many other approaches are conceivable and further research is needed to estimate their cost-effectiveness.[55]

8.3 Comparison with non-response bias

Endogenous sample selection is closely related to non-response bias, a long standing concern across fields in the social sciences (Dutz et al., 2021). It is similar to non-response because, like non-response, it leads to non-randomly missing data. At the same time, it is distinct from non-response and hitherto underappreciated because, unlike non-response, it is not directly observable in the data.

[55] For example, Tabe-Ojong et al. (2026) provide experimental evidence that eliciting the total number of household members before listing each member individually leads to 7% larger households.

How important is endogenous sample selection quantitatively compared to non-response? We compare the extent of endogenous sample selection to the extent of non-response to the man's questionnaire in our sample of DHS to quantify the relative contribution of each of these two margins to missing data from eligible men. Endogenous sample selection captures all those eligible men who were present in the interviewed households, but not recorded as such. Non-response captures all those eligible men recorded as such who did not complete the man's questionnaire. Across our sample of surveys, we find that endogenous sample selection and non-response account for missing data from eligible men in roughly equal shares. This implies that the issue of missing data is nearly twice as severe as reflected by non-response alone, with the average total share of eligible men without a completed man's questionnaire amounting to 17% rather than 9%, as previously assumed.[56]

9 Conclusion

This paper examines the production of human-collected data, crucial input to a wide range of data sources in the social sciences and beyond. We show that data collectors systematically screen out units that require disproportionate effort in collection based on ex ante-observable characteristics – either by omitting such units entirely or by manipulating their eligibility criteria. This data collector behaviour induces selection of units out of sample, and as a result biases statistics and analysis.

Exploiting the random assignment of effort cost across data collectors, we find that on average 8.8% of individuals eligible for individual interview are missing in two of the largest global household survey programs, the DHS and MICS. Within data collector, the expected higher effort cost of working under extreme temperatures is reflected in data collectors indeed excluding more eligible individuals on hot days.

Endogenous sample selection by data collectors is not random. Missing individuals disproportionately come from marginalised populations. They are more likely to be peripheral in their respective households, more likely to have never married, more likely to be less educated, sicker and poorer. This selection is consistent with data collectors screening out exactly those of the costly individuals whose exclusion is harder to detect due to ambiguity in their household affiliation.

[56] Appendix A.10.2 provides details on the extent of missing data from eligible men due endogenous sample selection and non-response in each survey and examines the resulting aggregate bias.

Non-random sample selection leads to bias in important aggregate statistics that the DHS and MICS were designed to track closely, such as fertility, child mortality and child marriage. In line with the profile of screened out individuals, we find fertility and child marriage to be over-reported, whereas child mortality is under-reported. The former matters for population projections and national policy, whereas the latter have implications for the monitoring of progress towards the SDGs, which relies on DHS/MICS data as a key source in many low-income countries.

In economic research, treatments may not only cause outcomes of interest to change, but also affect incentives for data collectors on the ground. For example, we show how prominent causes of economic development such as climate and institutions do not only cause economic development, as established in careful empirical work, but also cause differential sample selection. This finding provides a cautionary tale for the causal identification of phenomena using human-collected data.

Thus, our main contribution is to highlight the existence of a previously overlooked margin of sample selection: the selection of respondents into and out of sample by data collectors. We estimate that, in our sample, data collector manipulation to screen out respondents appears to be as large as non-response bias due to respondent self-selection. That bias from both sources is often additive causes further concern.

Since incentives for data collectors are endemic across a wide array of data collection instruments, our work has implications for human-collected data more broadly, beyond the DHS and MICS. Endogenous sample selection appears to affect various other household surveys across diverse contexts, as well as firm and farm censuses around the world. We leave a rigorous analysis of bias and consequences in the latter, where sample selection is anecdotally ripe, for future work.

Finally, our work points to a broader, underlying trade-off between information and bias due to endogenous sample selection that we expect to affect all human-collected data. In our context, higher workload in the form of additional questions asked, more modules and more time spent collecting information per respondent translates directly into more eligible respondents missing from the sample due to data collector manipulation. This information-bias trade-off invites reflection from policymakers and academics alike on the future of data collection.

References

Abay, K. A., Berhane, G., Hoddinott, J., & Tafere, K. (2022). Respondent Fatigue Reduces Dietary Diversity Scores Reported from Mobile Phone Surveys in Ethiopia during the COVID-19 Pandemic. *J Nutr.*, *152*(10), 2269–2276.

Abdel Ghany, J., Wilde, J., Dimitrova, A., Kashyap, R., & Muttarak, R. (2026). Temperature and sex ratios at birth. *PNAS*, *123*(8), e2422625123.

Adhvaryu, A., Kala, N., & Nyshadham, A. (2020). The light and the heat: Productivity co-benefits of energy-saving technology. *Rev Econ Stat*, *102*(4), 779–792.

Alesina, A., Giuliano, P., & Nunn, N. (2013). On the origins of gender roles: Women and the plough. *Q J Econ*, *128*(2), 469–530.

Ambler, K., Herskowitz, S., & Maredia, M. K. (2021). Are we done yet? Response fatigue and rural livelihoods. *J Dev Econ*, *153*, 102736.

Amirapu, A., & Gechter, M. (2020). Labor Regulations and the Cost of Corruption: Evidence from the Indian Firm Size Distribution. *Rev Econ Stat*, *102*(1), 34–48.

Asher, S., Campion, A., Gollin, D., & Novosad, P. (2024). *The Long-Run Development Impacts of Agricultural Productivity Gains: Evidence from Irrigation Canals in India* (CEPR Discussion Papers No. 17414). C.E.P.R. Discussion Papers.

Asher, S., & Novosad, P. (2017). Politics and local economic growth: Evidence from India. *AEJ: Applied Economics*, *9*(1), 229–73.

Asher, S., & Novosad, P. (2020). Rural roads and local economic development. *Am Econ Rev*, *110*(3), 797–823.

Axtell, R. L. (2001). Zipf distribution of us firm sizes. *Science*, *293*(5536), 1818–1820.

Bachi, R. (1951). The tendency to round off age returns: Measurement and correction. *Bulletin of the International Statistical Institute*, *33*(4), 195–222.

Becker, A. (2024). On the economic origins of concerns over women's chastity. *Rev Econ Stud*, *92*, 2303–2329.

Blakeslee, D., Dar, A., Fishman, R., Malik, S., Pellegrina, H. S., & Bagavathinathan, K. S. (2023). Irrigation and the spatial pattern of local economic development in india. *J Dev Econ*, *161*, 102997.

Blasius, J., & Thiessen, V. (2015). Should we trust survey data? assessing response simplification and data fabrication. *Social science research*, *52*, 479–93.

Blom, S., Ortiz-Bobea, A., & Hoddinott, J. (2022). Heat exposure and child nutrition: Evidence from west africa. *Journal of Env Econ and Management*, *115*, 102698.

Buera, F. J., Kaboski, J. P., & Shin, Y. (2020). The macroeconomics of microfinance. *Rev Econ Stud*, *88*(1), 126–161.

Burke, M., Gong, E., & Jones, K. (2015). Income shocks and hiv in africa. *The Economic Journal*, *125*(585), 1157–1189.

Burlig, F., & Preonas, L. (2024). Out of the Darkness and into the Light? Development Effects of Rural Electrification. *J Polit Econ*, *132*(9), 2937–2971.

Cai, X., Lu, Y., & Wang, J. (2018). The impact of temperature on manufacturing worker productivity: Evidence from personnel data. *J of Comp Econ*, *46*.

Chatterjee, S., & Vogl, T. (2018). Escaping Malthus: Economic Growth and Fertility Change in the Developing World. *Am Econ Rev*, *108*(6), 1440–67.

Chiplunkar, G., & Goldberg, P. K. (2024). Aggregate implications of barriers to female entrepreneurship. *Econometrica*, *92*(6), 1801–1835.

Corno, L., Hildebrandt, N., & Voena, A. (2020). Age of marriage, weather shocks, and the direction of marriage payments. *Econometrica*, *88*(3), 879–915.

Crespi, L. (1945). The Cheater Problem in Polling. *Public Opin Quart*, *9*(4), 431–445.

Cui, W., Cao, G., Park, J., Ouyang, Q., & Zhu, Y. (2013). Influence of indoor air temperature on human thermal comfort, motivation and performance. *Building and Environment*, *68*, 114–122.

Di Maio, M., & Fiala, N. (2019). Be Wary of Those Who Ask: A Randomized Experiment on the Size and Determinants of the Enumerator Effect. *The World Bank Economic Review*, *34*(3), 654–669.

Dillon, A., & Mensah, E. (2024). Respondent biases in agricultural household surveys. *J Dev Econ*, *166*, 103198.

Dimitrova, A. (2021). Seasonal droughts and the risk of childhood undernutrition in ethiopia. *World Development*, *141*, 105417.

Dimitrova, A., Dimitrova, A., Mengel, M., Gasparrini, A., Lotze-Campen, H., & Gabrysch, S. (2024). Temperature-related neonatal deaths attributable to climate change in 29 lmics. *Nature Communications*, *15*(1), 5504.

Dupas, P., Falezan, C., Mabeu, M. C., & Rossi, P. (2023). *Long-run impacts of forced labor migration on fertility behaviors: Evidence from colonial West Africa* (Working Paper No. 31993).

Dutz, D., Huitfeldt, I., Lacouture, S., Mogstad, M., Torgovitsky, A., & van Dijk, W. (2021). *Selection in Surveys: Using Randomized Incentives to Detect and Account for Nonresponse Bias* (Working Paper No. 29549). National Bureau of Economic Research.

Eckman, S., & Koch, A. (2019). Interviewer involvement in sample selection shapes the relationship between response rates and data quality. *Public Opinion Q*, *83*, 313–337.

Epstein, A., Bendavid, E., Nash, D., Charlebois, E. D., & Weiser, S. D. (2020). Drought and ipv towards women in 19 countries in sub-saharan africa during 2011-2018: A population-based study. *PLoS Medicine*, *17*(3), e1003064.

Figueiredo Walter, T., Moneke, N., & Radu, A. (forthcoming). The missing poor. *American Economic Review: Insights*.

Finn, A., & Ranchhod, V. (2015). Genuine Fakes: The Prevalence and Implications of Data Fabrication in a Large South African Survey. *The World Bank Economic Review*, *31*(1), 129–157.

Gerard, F., Rokkanen, M., & Rothe, C. (2020). Bounds on treatment effects in rd designs with a manipulated running variable. *Quant Econ*, *11*(3).

Geruso, M., & Spears, D. (2018). *Heat, humidity, and infant mortality in the developing world* (tech. rep. No. 24870). National Bureau of Economic Research.

Gray, C., & Thiede, B. C. (2024). Temperature anomalies undermine the health of reproductive-age women in low- and middle-income countries. *Proceedings of the National Academy of Sciences*, *121*(11), e2311567121.

He, C., Zhu, Y., Guo, Y., Bell, M. L., Filippi, V., Brimicombe, C., Chen, R., & Kan, H. (2025). Rainfall variability and under-five child mortality in 59 low- and middle-income countries. *Nature Water*, *3*(8), 881–889.

Hernández-Pérez, R., Angulo-Brown, F., & Tun, D. (2006). Company size distribution for developing countries. *Physica A: Statistical Mechanics*, *359*, 607–618.

Honoré, B. E., & Hu, L. (2020). Selection without exclusion. *Econometrica*, *88*(3), 1007–1029.

Horrigan, M., Moore, W., Pedlow, S., & Wolter, K. (1999). Undercoverage in a large national screening survey for youths. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 570–75.

ICF. (1982-2022). *Demographic and health surveys (various)* [Datasets used: Various]. Rockville, Maryland, ICF.

Jeong, D., Aggarwal, S., Robinson, J., Kumar, N., Spearot, A., & Park, D. S. (2023). Exhaustive or exhausting? Evidence on respondent fatigue in long surveys. *J Dev Econ*, *161*, 102992.

Kim, H. B., Kim, S., & Kim, T. T. (2020). The role of career and wage incentives in labor productivity: Evidence from a two-stage field experiment in malawi. *Rev Econ Stat*, *102*(5), 839–851.

Kosyakova, Y., Skopek, J., & Eckman, S. (2014). Do interviewers manipulate responses to filter questions? evidence from a multilevel approach. *International Journal of Public Opinion Research*, *27*(3), 417–431.

Le, K., & Nguyen, M. (2022). Droughts and child health in Bangladesh. *Plos one*, *17*(3), e0265617.

Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Rev Econ Stud*, *76*(3), 1071–1102.

LoPalo, M. (2023). Temperature, Worker Productivity, and Adaptation: Evidence from Survey Data Production. *AEJ: Applied Economics*, *15*(1), 192–229.

Lowes, S., & Montero, E. (2021a). Concessions, violence, and indirect rule: Evidence from the congo free state. *Q J Econ*, *136*(4), 2047–2091.

Lowes, S., & Montero, E. (2021b). The legacy of colonial medicine in central africa. *Am Econ Rev*, *111*(4), 1284–1314.

Marckwardt, A. M., & Rutstein, S. O. (1996). *Accuracy of dhs-ii demographic data: Gains and losses in comparison with earlier surveys* (tech. rep.). Calverton, Maryland, USA, Macro International.

Masselus, L., & Fiala, N. (2024). Whom to ask? Testing respondent effects in household surveys. *J Dev Econ*, *168*, 103265.

Menold, N., Landrock, U., Winker, P., Pellner, N., & Kemper, C. J. (2018). The impact of payment and respondents' participation on interviewers' accuracy in face-to-face surveys. *Field Methods*, *30*(4), 295–311.

Meyer, B. D., Mok, W. K. C., & Sullivan, J. X. (2015). Household Surveys in Crisis. *Journal of Economic Perspectives*, *29*(4), 199–226.

Ministry of Statistics and Programme Implementation, India. (2005). *Instruction Manual Fifth Economic Census (2005).*

Moore, W., Pedlow, S., Krishnamurty, P., & Wolter, K. (2000). National longitudinal survey of youth 1997. *National Opinion Research Center, Chicago, IL*, *254*.

Moscona, J., Nunn, N., & Robinson, J. A. (2020). Segmentary lineage organization and conflict in sub-saharan africa. *Econometrica*, *88*(5), 1999–2036.

Muralidharan, K., Niehaus, P., & Sukhtankar, S. (2023). General equilibrium effects of (improving) public employment programs: Experimental evidence from india. *Econometrica*, *91*(4), 1261–1295.

Myers, R. J. (1940). Errors and bias in the reporting of ages in census data. *Transactions of the Actuarial Society of America*, *41*, 395–415.

Nagata, J. M., Epstein, A., Ganson, K. T., Benmarhnia, T., & Weiser, S. D. (2021). Drought and child vaccination coverage in 22 countries in sub-saharan africa: A retrospective analysis of national survey data from 2011 to 2019. *PLoS medicine*, *18*(9), e1003678.

Nolan, L., Lucas, R., Choi, Y., Short Fabic, M., & Adetunji, J. (2017). The contribution of demographic and health survey data to population and health policymaking: Evidence from three developing countries. *Etude de la population africaine*, *31*, 3394–3407.

Norling, J. (2022). Fertility following natural disasters and epidemics in africa. *The World Bank Economic Review*, *36*(4), 955–971.

Nunn, N. (2008). The long-term effects of Africa's slave trades. *Q J Econ*, *123*(1), 139–176.

Nunn, N., & Wantchekon, L. (2011). The slave trade and the origins of mistrust in africa. *Am Econ Rev*, *101*(7), 3221–52.

Pilcher, J. J., Nadler, E., & Busch, C. (2002). Effects of hot and cold temperature exposure on performance: A meta-analytic review. *Ergonomics*, *45*, 682–98.

Pullum, T. W. (2006). *An assessment of age and date reporting in the dhs surveys, 1985-2003* (tech. rep.). Calverton, Maryland, USA, Macro International.

Rossi, P. (2018). Strategic Choices in Polygamous Households: Theory and Evidence from Senegal. *Rev Econ Stud*, *86*(3), 1332–1370.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592.

Ruggles, S., Cleveland, L., Lovaton, R., Sarkar, S., Sobek, M., Burk, D., Ehrlich, D., Heimann, Q., & Lee, J. (2024). *Integrated public use microdata series, international: Version 7.5.* Minneapolis, MN, IPUMS.

Sen, R. (2024). Supervision at work: Evidence from a field experiment. *LMU mimeo.*

Seppanen, O., Fisk, W., & Lei, Q. H. (2005). Effect of temperature on task performance in office environment.

Short Fabic, M., Choi, Y., & Bird, S. (2012). A systematic review of demographic and health surveys: Data availability and utilization for research. *Bulletin of the World Health Organization*, *90*, 604–12.

Somanathan, E., Somanathan, R., Sudarshan, A., & Tewari, M. (2021). The impact of temperature on productivity and labor supply: Evidence from indian manufacturing. *J Polit Econ*, *129*(6), 1797–1827.

Sustainable Development Solutions Network. (2015). Data for development: A needs assessment for SDG monitoring and statistical capacity development.

Tabe-Ojong, M. P. J., Abay, K. A., Abate, G. T., & Machio, P. M. (2026). When the going gets tough: Experimental evidence of respondent fatigue in household surveys. *Journal of African Economies*, *35*(2), ejag001.

Thiede, B. C. (2014). Rainfall shocks and within-community wealth inequality: Evidence from rural Ethiopia. *World Development*, *64*, 181–193.

Thiede, B. C., & Strube, J. (2020). Climate variability and child nutrition: Findings from sub-Saharan Africa. *Global Environmental Change*, *65*, 102192.

UNICEF. (2000-2022). *Multiple indicator cluster surveys (mics)* [Datasets used: Various]. New York, USA, UNICEF [Distributor].

Vogl, T. S. (2016). Differential Fertility, Human Capital, and Development. *Rev Econ Stud*, *83*(1), 365–401.

West, B., & Blom, A. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, *5*, 175–211.

Zipfel, C. (forthcoming). The demand side of Africa's demographic transition: Desired fertility, wealth, and jobs. *Journal of Human Resources*.

A Online appendix

A.1 Use of DHS and MICS for policy and research

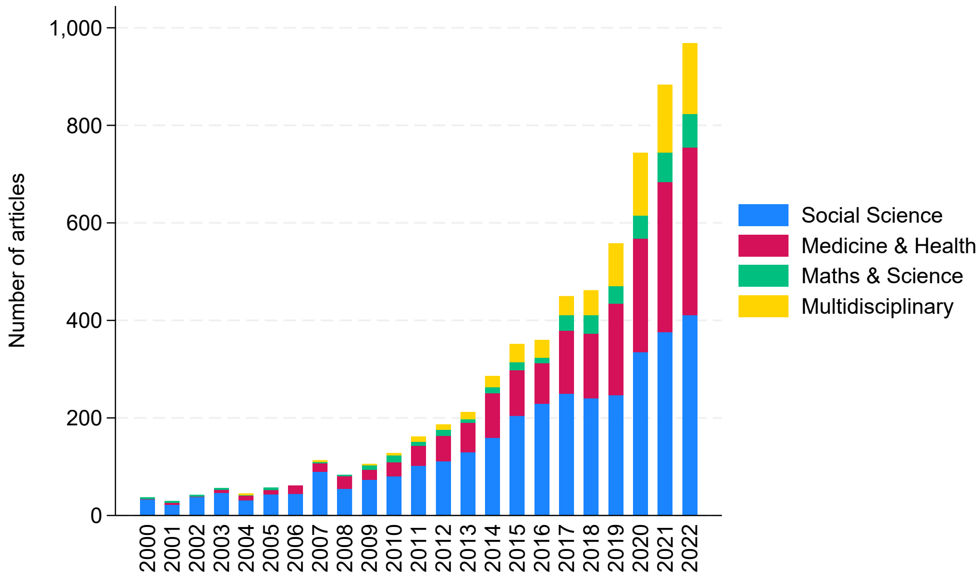


Figure A1: Use of DHS and MICS in research over time

This figure displays the count of journal articles that refer to the DHS or MICS in their title or abstract across disciplines over time. Counts were generated from the Web of Science database using keyword search. The set of journals is restricted to those that formed part of the Essential Science Indicator journal master list as of June 2024.

A.2 Sample selection and data construction

A.2.1 Selection of surveys

The main criterion for the inclusion of a survey into our main sample is the administration of a man's questionnaire in a randomly selected subset of households. Additionally, we restrict our sample to nationally representative surveys. This enables us to examine implications of endogenous sample selection for national statistics.

We identify relevant surveys from the official survey reports published on the DHS and MICS websites. To this end, we read more than 800 reports in five different languages and extract information on all survey components that were randomly varied across households, most importantly the man's questionnaire, biomarker collection and the domestic violence module. The combination of the information from the reports and the microdata allows us to understand the underlying randomisation in detail. In particular, we pay close attention to the manner in which different randomised survey features were either bundled or cross-randomised and the respective treatment probabilities.

Among all 236 surveys that satisfy our criteria, we exclude 55 because they do not lend themselves to our analysis due to differences in survey design or data issues. All excluded surveys and the respective reasons for exclusion are listed in Table A1. First, we exclude 28 surveys that administered additional survey features, such as biomarker collection among children, in control households (without a man's questionnaire). In these cases, differences in outcomes between treatment and control households cannot be attributed solely to the man's questionnaire. Second, we exclude 13 surveys in which eligibility for the man's questionnaire is conditional on marital status. Selection into individual questionnaires in these surveys is not comparable to selection in included surveys and thus results would not be directly comparable. Moreover, the resulting samples are not nationally representative. Third, we exclude 9 MICS due to data issues. For 6 MICS in which sampling is stratified by enumeration area and the presence of children in the household, we do not observe the latter stratification variable in the microdata. Thus we cannot control for stratum fixed effects. For 3 MICS, we are not able to merge the individual- and household-level microdata source files because identifiers do not match across files. Fourth, 3 DHS are excluded because their man's questionnaire does not have an upper age limit, preventing us from defining a comparable group of ineligible men in these surveys. Finally, one DHS is excluded because treatment was randomised across enumeration areas rather than across households within enumeration areas, making comparisons with other surveys difficult. One MICS is excluded due to contradicting information about treatment assignment in the survey report and the microdata.

Table A1: Excluded surveys with randomly assigned man's questionnaire

Reason for exclusion	**Excluded surveys**	**Total**
Additional survey features administered in control households (without man's questionnaire) that were not implemented in treatment households (with man's questionnaire)	AGO DHS 2015; BEN DHS 2017; CIV MICS 2016; CMR DHS 2004, 2011, 2018; COD DHS 2013; COD MICS 2017; COG DHS 2011; COM MICS 2022; DOM DHS 2002; GIN DHS 2012; JOR DHS 2017; KAZ DHS 1999; KHM DHS 2005, 2021; MDG DHS 2021; MDG MICS 2018; MOZ DHS 2011; MRT DHS 2019; NPL DHS 2022; RWA DHS 2014, 2019; SEN DHS 2018, 2018, 2019; TCD DHS 2014; TLS DHS 2016	28
Eligibility for man's questionnaire conditional on marital status	AFG DHS 2015; BGD DHS 1996, 1999, 2007, 2011; IDN DHS 2002, 2007, 2012, 2017; MDV DHS 2009; NPL DHS 2001; PAK DHS 2012, 2017	13
Randomisation of man's questionnaire stratified by presence of children at household listing stage, but variable not available in microdata	BLR MICS 2012, 2019; GUY MICS 2014; MNE MICS 2013, 2018; UKR MICS 2012	6
No upper age limit for man's questionnaire eligibility	BFA DHS 1993, MAR DHS 1992, SEN DHS 1992	3
Individual identifiers do not match across microdata source files	STP MICS 2019; SWZ MICS 2014; TGO MICS 2017	3
Random assignment of man's questionnaire across clusters rather than households within clusters	GHA DHS 1993	1
Contradicting info about assignment of man's questionnaire in survey report and microdata	KAZ MICS 2010	1

Table A2: DHS and MICS with randomly assigned man's questionnaire

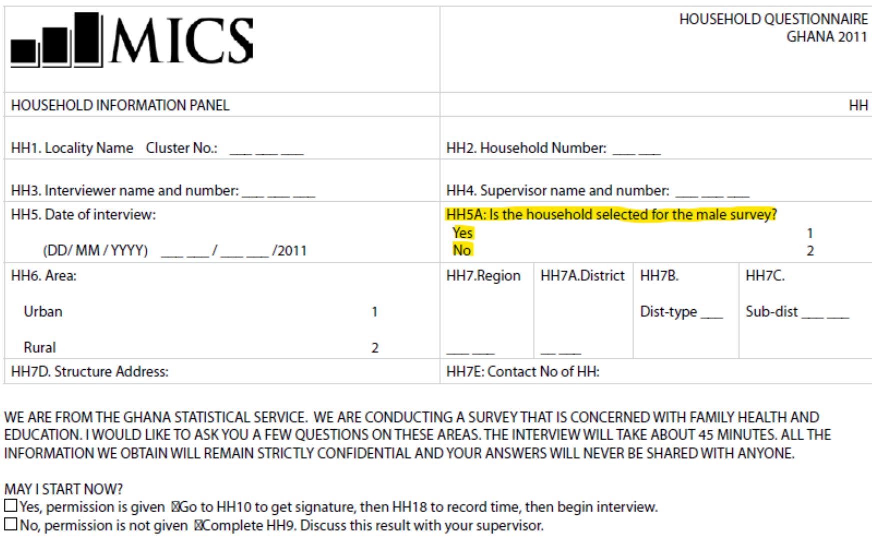
Country code	Country name	DHS	MICS
ALB	Albania	2008, 2017	NA
ARM	Armenia	2000, 2005, 2010, 2015	NA
AZE	Azerbaijan	2006	NA
BDI	Burundi	2010, 2016	NA
BEN	Benin	1996, 2001, 2006, 2011	2014
BFA	Burkina Faso	1998, 2003, 2010, 2021	NA
BGD	Bangladesh	2004	NA
BOL	Bolivia	1998, 2003, 2008	NA
BRA	Brazil	1996	NA
CAF	Central African Republic	1994	2006, 2010, 2018
CIV	Côte d'Ivoire	1994, 1998, 2011, 2021	NA
CMR	Cameroon	1998	2014
COD	Congo - Kinshasa	2007	NA
COG	Congo - Brazzaville	2005	2014
COM	Comoros	1996, 2012	NA
CUB	Cuba	NA	2014, 2019
ETH	Ethiopia	2000, 2005	NA
FJI	Fiji	NA	2021
GAB	Gabon	2000, 2012, 2019	NA
GEO	Georgia	NA	2018
GHA	Ghana	1998, 2008, 2014	2006, 2011, 2017
GIN	Guinea	1999, 2005, 2018	NA
GMB	Gambia	2013, 2019	2018
GNB	Guinea-Bissau	NA	2014, 2018
GTM	Guatemala	2014	NA
HND	Honduras	2011	2019
HTI	Haiti	1994, 2000, 2005, 2012	NA
IND	India	2005, 2015, 2019	NA
KEN	Kenya	1993, 1998, 2003, 2008, 2014, 2022	NA
KGZ	Kyrgyzstan	2012	NA
KHM	Cambodia	2010, 2014	NA
KIR	Kiribati	NA	2018
LAO	Laos	NA	2017
LBR	Liberia	2013, 2019	NA
LSO	Lesotho	2004, 2009, 2014	2018
MDA	Moldova	2005	2012
MDG	Madagascar	2003, 2008	NA
MLI	Mali	1995, 2001, 2006, 2012, 2018	2015
MMR	Myanmar (Burma)	2015	NA
MNG	Mongolia	NA	2013, 2018
MOZ	Mozambique	1997, 2003	NA
MRT	Mauritania	NA	2007, 2015
MWI	Malawi	1992, 2000, 2004, 2010, 2015	2006, 2013, 2019
NAM	Namibia	2000, 2006, 2013	NA
NER	Niger	1998, 2006, 2012	NA
NGA	Nigeria	2003, 2008, 2013, 2018	NA
NIC	Nicaragua	1998	NA
NPL	Nepal	2006, 2011, 2016	2019
PER	Peru	1996	NA
PHL	Philippines	2003	NA
PNG	Papua New Guinea	2016	NA
RWA	Rwanda	2000, 2005, 2010	NA
SEN	Senegal	2005, 2010, 2014, 2015, 2016	NA
SLE	Sierra Leone	2008, 2013, 2019	2017
SUR	Suriname	NA	2018
TCA	Turks and Caicos Islands	NA	2019
TCD	Chad	1996, 2004	2019
TGO	Togo	1998, 2013	2010
THA	Thailand	NA	2019, 2022
TLS	Timor-Leste	2009	NA
TON	Tonga	NA	2019
TUN	Tunisia	NA	2018
TUV	Tuvalu	NA	2019
TZA	Tanzania	1991, 1996, 2004, 2010, 2015, 2022	NA
UGA	Uganda	1995, 2000, 2006, 2011, 2016	NA
UKR	Ukraine	2007	NA
UZB	Uzbekistan	2002	NA
VNM	Vietnam	NA	2020
WSM	Samoa	NA	2019
XKX	Republic of Kosovo	NA	2013, 2019
ZAF	South Africa	2016	NA
ZMB	Zambia	1996, 2001	NA
ZWE	Zimbabwe	1994, 1999	2014, 2019

Table A3: MICS/DHS-Population Census pairs

Country	Survey	Survey Year	PHC Year	Source	Statistical Office
AGO	DHS	2015	2014	IPUMS	Instituto Nacional de Estatistica
BEN	DHS	2001	2002	IPUMS	National Institute of Statistics and Economic Analysis
BEN	DHS	2011	2013	IPUMS	National Institute of Statistics and Economic Analysis
BEN	MICS	2014	2013	IPUMS	National Institute of Statistics and Economic Analysis
BFA	DHS	1998	1996	IPUMS	National Institute of Statistics and Demography
BFA	MICS	2006	2006	IPUMS	National Institute of Statistics and Demography
BOL	DHS	1994	1992	IPUMS	National Institute of Statistics
BOL	DHS	2003	2001	IPUMS	National Institute of Statistics
CIV	DHS	1998	1998	IPUMS	Agence National de la Statistique
CMR	DHS	2004	2005	IPUMS	Central Bureau of Census and Population Studies
CMR	MICS	2006	2005	IPUMS	Central Bureau of Census and Population Studies
CRI	MICS	2011	2011	IPUMS	National Institute of Statistics and Censuses
CUB	MICS	2010	2012	IPUMS	Office of National Statistics
CUB	MICS	2014	2012	IPUMS	Office of National Statistics
DOM	MICS	2000	2002	IPUMS	National Statistics Office
GHA	DHS	1998	2000	IPUMS	Ghana Statistical Services
GHA	DHS	2008	2010	IPUMS	Ghana Statistical Services
GIN	DHS	2012	2014	IPUMS	National Statistics Directorate
IDN	MICS	2000	2000	IPUMS	Statistics Indonesia
KEN	DHS	1989	1989	IPUMS	National Bureau of Statistics
KEN	DHS	1998	1999	IPUMS	National Bureau of Statistics
KEN	DHS	2008	2009	IPUMS	National Bureau of Statistics
KHM	DHS	2000	1998	IPUMS	National Institute of Statistics
KHM	DHS	2005	2004	IPUMS	National Institute of Statistics
KHM	DHS	2010	2008	IPUMS	National Institute of Statistics
KHM	DHS	2014	2013	IPUMS	National Institute of Statistics
KHM	DHS	2021	2019	IPUMS	National Institute of Statistics
LAO	MICS	2006	2005	IPUMS	Statistics Bureau
LAO	MICS	2017	2015	IPUMS	Statistics Bureau
LBR	DHS	2007	2008	IPUMS	Institute of Statistics and Geo-Information Systems
LBR	DHS	2009	2008	IPUMS	Institute of Statistics and Geo-Information Systems
LSO	DHS	2004	2006	IPUMS	Bureau of Statistics
MAR	DHS	1992	1994	IPUMS	High Commission of Planning
MAR	DHS	2003	2004	IPUMS	High Commission of Planning
MEX	MICS	2015	2015	IPUMS	National Institute of Statistics, Geography, and Informatics
MMR	DHS	2015	2014	IPUMS	Central Statistical Organization
MNG	MICS	2010	2010	NSO	National Statistical Office
MOZ	DHS	1997	1997	IPUMS	National Institute of Statistics
MOZ	MICS	2008	2007	IPUMS	National Institute of Statistics
MOZ	DHS	2009	2007	IPUMS	National Institute of Statistics
MWI	DHS	1996	1998	IPUMS	National Statistical Office
MWI	DHS	2000	1998	IPUMS	National Statistical Office
MWI	MICS	2006	2008	IPUMS	National Statistical Office
MWI	DHS	2010	2008	IPUMS	National Statistical Office
NER	DHS	2012	2012	NSO	National Institute of Statistics
PER	DHS	1991	1993	IPUMS	National Institute of Statistics and Informatics
PER	DHS	2009	2007	IPUMS	National Institute of Statistics and Informatics
PRY	DHS	1990	1992	IPUMS	General Directorate of Statistics, Surveys, and Censuses
RWA	DHS	1992	1991	IPUMS	National Institute of Statistics
RWA	DHS	2000	2002	IPUMS	National Institute of Statistics
RWA	MICS	2000	2002	IPUMS	National Institute of Statistics
RWA	DHS	2010	2012	IPUMS	National Institute of Statistics
RWA	DHS	2013	2012	IPUMS	National Institute of Statistics
RWA	DHS	2014	2012	IPUMS	National Institute of Statistics
SEN	DHS	2012	2013	IPUMS	National Agency of Statistics and Demography
SEN	DHS	2014	2013	IPUMS	National Agency of Statistics and Demography
SEN	DHS	2015	2013	IPUMS	National Agency of Statistics and Demography
SLE	DHS	2013	2015	IPUMS	Statistics Sierra Leone
SLE	DHS	2016	2015	IPUMS	Statistics Sierra Leone
TGO	MICS	2010	2010	IPUMS	National Institute of Statistics (INSEED)
TTO	MICS	2011	2011	IPUMS	Central Statistical Office
TZA	DHS	2003	2002	IPUMS	National Bureau of Statistics
TZA	DHS	2004	2002	IPUMS	National Bureau of Statistics
TZA	DHS	2010	2012	IPUMS	National Bureau of Statistics
TZA	DHS	2011	2012	IPUMS	National Bureau of Statistics
UGA	DHS	2000	2002	IPUMS	Bureau of Statistics
UGA	DHS	2014	2014	IPUMS	Bureau of Statistics
UGA	DHS	2016	2014	IPUMS	Bureau of Statistics
URY	MICS	2012	2011	IPUMS	National Institute of Statistics
VEN	MICS	2000	2001	IPUMS	National Institute of Statistics
VNM	MICS	2010	2009	IPUMS	General Statistics Office
VNM	MICS	2020	2019	IPUMS	General Statistics Office
ZMB	DHS	1992	1990	IPUMS	Central Statistical Office
ZMB	DHS	2001	2000	IPUMS	Central Statistical Office
ZWE	DHS	2010	2012	IPUMS	Central Statistical Office

A.2.2 Eligibility for individual questionnaires

To determine the age thresholds for the eligibility of household members for individual questionnaires, we systematically extract information on the age thresholds from the official survey reports and questionnaires for all surveys in our sample. Subsequently, we verify the consistency of the microdata with these thresholds.



MICS	HOUSEHOLD QUESTIONNAIRE GHANA 2011			
HOUSEHOLD INFORMATION PANEL	HH			
HH1. Locality Name Cluster No.: ___ ___ ___	HH2. Household Number: ___ ___			
HH3. Interviewer name and number: ___ ___ ___	HH4. Supervisor name and number: ___ ___ ___			
HH5. Date of interview: (DD/ MM / YYYY) ___ ___ / ___ ___ /2011	HH5A: Is the household selected for the male survey? Yes 1 No 2			
HH6. Area: Urban 1 Rural 2	HH7.Region ___ ___	HH7A.District ___ ___	HH7B. Dist-type ___	HH7C. Sub-dist ___ ___
HH7D. Structure Address:	HH7E: Contact No of HH:			

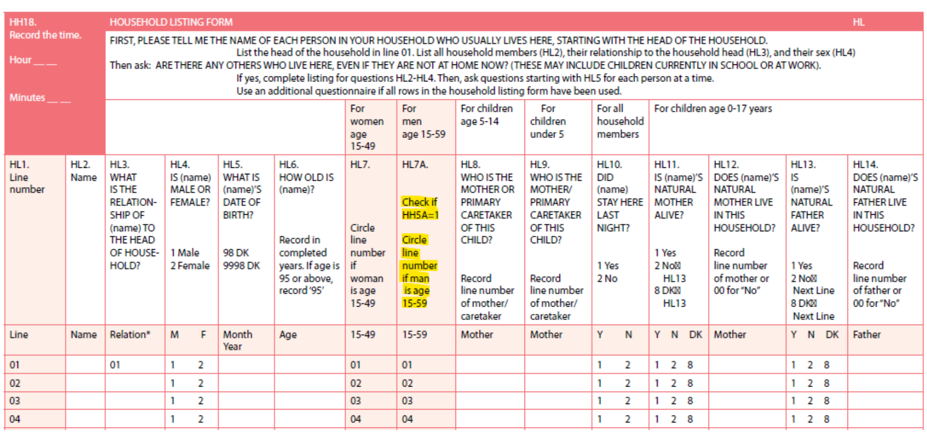
WE ARE FROM THE GHANA STATISTICAL SERVICE. WE ARE CONDUCTING A SURVEY THAT IS CONCERNED WITH FAMILY HEALTH AND EDUCATION. I WOULD LIKE TO ASK YOU A FEW QUESTIONS ON THESE AREAS. THE INTERVIEW WILL TAKE ABOUT 45 MINUTES. ALL THE INFORMATION WE OBTAIN WILL REMAIN STRICTLY CONFIDENTIAL AND YOUR ANSWERS WILL NEVER BE SHARED WITH ANYONE.

MAY I START NOW?

☐ Yes, permission is given Go to HH10 to get signature, then HH18 to record time, then begin interview.

☐ No, permission is not given Complete HH9. Discuss this result with your supervisor.

Figure A2: MICS, Ghana 2011: First page of household questionnaire



HH18. Record the time. Hour ___ ___ Minutes ___ ___

HOUSEHOLD LISTING FORM — HL

FIRST, PLEASE TELL ME THE NAME OF EACH PERSON IN YOUR HOUSEHOLD WHO USUALLY LIVES HERE, STARTING WITH THE HEAD OF THE HOUSEHOLD.
List the head of the household in line 01. List all household members (HL2), their relationship to the household head (HL3), and their sex (HL4)
Then ask: ARE THERE ANY OTHERS WHO LIVE HERE, EVEN IF THEY ARE NOT AT HOME NOW? (THESE MAY INCLUDE CHILDREN CURRENTLY IN SCHOOL OR AT WORK).
If yes, complete listing for questions HL2-HL4. Then, ask questions starting with HL5 for each person at a time.
Use an additional questionnaire if all rows in the household listing form have been used.

						For women age 15-49	For men age 15-59	For children age 5-14	For children under 5	For all household members	For children age 0-17 years			
HL1. Line number	HL2. Name	HL3. WHAT IS THE RELATION-SHIP OF (name) TO THE HEAD OF HOUSE-HOLD?	HL4. IS (name) MALE OR FEMALE? 1 Male 2 Female	HL5. WHAT IS (name)'S DATE OF BIRTH? 98 DK 9998 DK	HL6. HOW OLD IS (name)? Record in completed years. If age is 95 or above, record '95'	HL7. Circle line number if woman is age 15-49	HL7A. Check if HH5A=1 Circle line number if man is age 15-59	HL8. WHO IS THE MOTHER OR PRIMARY CARETAKER OF THIS CHILD? Record line number of mother/ caretaker	HL9. WHO IS THE MOTHER/ PRIMARY CARETAKER OF THIS CHILD? Record line number of mother/ caretaker	HL10. DID (name) STAY HERE LAST NIGHT? 1 Yes 2 No	HL11. IS (name)'S NATURAL MOTHER ALIVE? 1 Yes 2 No HL13 8 DK HL13	HL12. DOES (name)'S NATURAL MOTHER LIVE IN THIS HOUSEHOLD? Record line number of mother or 00 for "No"	HL13. IS (name)'S NATURAL FATHER ALIVE? 1 Yes 2 No Next Line 8 DK Next Line	HL14. DOES (name)'S NATURAL FATHER LIVE IN THIS HOUSEHOLD? Record line number of father or 00 for "No"
Line	Name	Relation*	M F	Month Year	Age	15-49	15-59	Mother	Mother	Y N	Y N DK	Mother	Y N DK	Father
01		01	1 2			01	01			1 2	1 2 8		1 2 8	
02			1 2			02	02			1 2	1 2 8		1 2 8	
03			1 2			03	03			1 2	1 2 8		1 2 8	
04			1 2			04	04			1 2	1 2 8		1 2 8	

Figure A3: MICS, Ghana 2011: Household roster

A.2.3 Data collector effort cost

We construct two proxies of the effort cost associated with household members of a given sex and age.

Questions listed. For all surveys in our respective samples, we count the total number of questions contained in the household roster (individual-level questions in the household questionnaire), the man's questionnaire and the woman's questionnaire. We proxy the effort cost associated with a (wo)man of eligible age with the sum of the number of questions in roster and individual questionnaire. The effort cost of ineligible household members is measured by the number of questions in the household roster.

We count questions as follows. We follow the numbering of questions in the official questionnaires and do not count sub-questions. For example, questions 32, 32A and 32B are counted as single question. Note that a small set of questions may be repeated multiple times for the same respondent. For example, women in recent DHS are asked several questions about each birth they have ever given. Independent of the number of births a woman has given, we only count each of these questions once. To ensure accurate counting, we conduct two independent counts for a sub-sample of 33 surveys. Reassuringly, we find a correlation coefficient of above 0.99 between counts, with a mean absolute deviation of less than 1%.

When counting questions in population and housing censuses, we differentiate between individual-level questions asked to women of fertile age (typically 12 years and older) and all other individual level questions. We think of the former questions as the equivalent of the woman's questionnaire and the latter questions as the equivalent of the household roster in the DHS and the MICS.

Questions asked. The number of questions asked to a given respondent is usually smaller than the total number of questions contained in questionnaires. This is because certain subsets of questions are only asked to respondents with specific characteristics. For example, in the MICS only women of eligible age who have ever given birth are asked about their birth history. To count the number of questions actually answered by each respondent, we manually match each question in the questionnaire with the corresponding variable in the microdata. In the MICS, there is a one-to-one link between questions listed in the questionnaire and variables in the dataset. Moreover, variable names in the microdata follow the question numbering in questionnaire, facilitating the matching. In the DHS and PHC, this is not the case. IPUMS source variables have descriptive variable names that help with matching. DHS matching relies on variable labels and tabulations as variable names cannot be used due to DHS recoding process that names variables using standardised codes (e.g., hv104). Given the large number of questions in the DHS, the resulting matching process is tedious and time-consuming (5-8 hours per survey). Therefore, we only conduct this exercise for a subset of DHS (31) while we complete it for all MICS in our sample.

In each of the three data sources, we ensure a variable is coded as missing if and only if the matched question was not asked about a given individual. Subsequently, we count the number of non-missing entries across all variables for each household member. To obtain a measure of the effort cost associated with a given sex and age, we average the number of questions asked within sex-age cells.

A.2.4 Outcome variables

Ever married. We define having ever been married in a broad sense. In line with most surveys in our sample, we count all individuals that are married, living with a partner, separated, divorced or widowed as ever married. Information on the marital status is collected through different questionnaires in the surveys we work with. In the MICS, marital status is asked in the individual questionnaire, not in the household roster. The DHS initially operated in the same way, but gradually moved to systematically including a question about marital status in the household roster. While the roster only features a question on marital status in some of the DHS conducted prior to 2012, it includes such a question for all surveys in our sample conducted thereafter. Therefore, we observe the marital status of men in control households in all DHS conducted post 2012 and in a subset of DHS conducted earlier.

Close relationship to household head. Nearly all censuses and surveys in our samples elicit information on the relationship of household members to the household head. The set of answer options varies greatly across surveys and censuses, however. To harmonise the information, we create an indicator variable that equals to 1 if a household member is closely related to the head of the household and zero otherwise. We define children, spouse(s), parents, parents-in-law and grandchildren as closely related to the head, and other relatives (e.g., uncles, nephews) and unrelated household members (e.g., domestic workers) as distantly related.

Years of schooling. Information on years of schooling is readily available in harmonised form in DHS and IPUMS-International census data. In the MICS and non-IPUMS censuses, we harmonise this information ourselves, combining information on the highest level and grade of education completed with the structure of the education system at the time of the survey. Note that we only consider formal education.

Number of biological children in the household. Most surveys in our sample include a module on the survival of parents in the household roster. For all children aged 17 and below, this module asks whether the biological mother and father are alive, and if so whether they live in the household. If the answer to both of these questions is affirmative, their line number is recorded. We measure the number of biological children each household member lives with by counting the number of children in the household for which they are indicated as the parent.

Children ever born. This variable captures the total number of children ever born alive to a woman. It is top-coded in some censuses. To ensure comparability with matched surveys, we apply the same top-coding to the matched surveys.

Poor. This variable is an indicator variable that takes value one if an individual does not possess all of the following three items: shoes, clothes and a blanket. Otherwise, it takes value zero. The underlying information is elicited as part of a module on basic needs of children between ages 5 and 17. The module is included in the following DHS in our sample: NGA 2008, HTI 2005, MWI 2010, NAM 2006, UGA 2006.

Sick. This variable is an indicator variable that takes value one if an individual has been very sick for at least 3 months during the past 12 months. Otherwise, it takes value zero. The underlying information is elicited as part of a module on chronic disease which is limited to adults between the ages of 18 and 59 in most surveys. The module is included in the following DHS in our sample: MLI 2006, NER 2006, RWA 2005, SEN 2005, UGA 2006, NGA 2008, HTI 2005, COD 2007 and MWI 2004.

Disabled. This variable is an indicator variable that takes value one if an individual is classified as disabled and zero otherwise. It is available in the household roster of the following DHS in our sample: BOL 1998, GHA 1998, GMB 2013, KHM 2014, MLI 2018, TZA 2022, UGA 2016 and ZAF 2016. We define a person as disabled if they suffer from at least one form of disability (e.g., blind, deaf, etc.). In surveys where the extent to which individuals have difficulties with certain activities (e.g., seeing, hearing, moving, etc.) is elicited, we consider individuals as disabled if they cannot do at least one activity at all or they can only do it with a lot of difficulty.

A.2.5 Survey characteristics

We manually extract information on survey implementation from every report for each of the 181 surveys in our sample, systematically coding up the below variables.

Field check tables. We determine if field check tables were used during survey implementation. These tables are sometimes also referred to as quality control tables and contain descriptive statistics of key indicators. They are produced regularly throughout the fieldwork period and are used to provide feedback to survey teams.

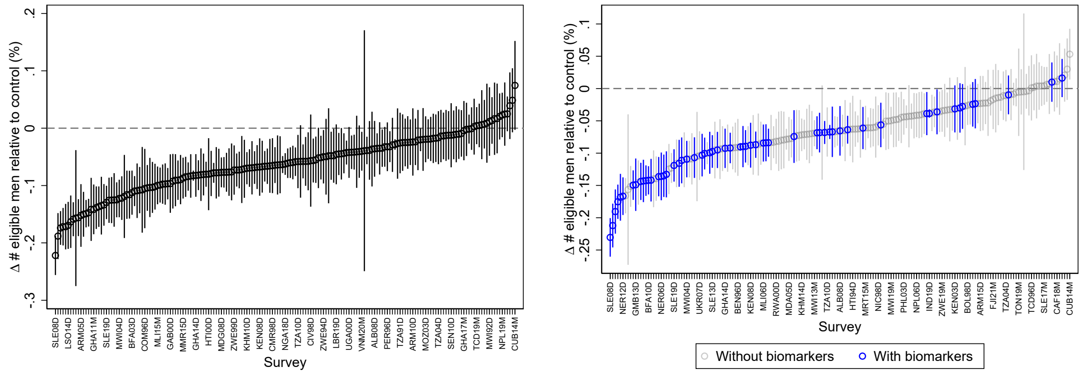
Mandatory re-interviewing. We identify surveys that conduct mandatory re-interviewing. In this case, typically two sets of households are re-interviewed: first, a random subset of households in each enumeration area and second, all households which have been identified as outliers along key survey dimensions.

Use of tablets. We differentiate between surveys that use paper and tablet questionnaires. In the former case, responses are recorded on paper and later entered into computers. In the latter case, responses are directly recorded on tablets and later transmitted to a central server.

A.3 Missing men: additional results

A.3.1 Robustness and heterogeneous effects

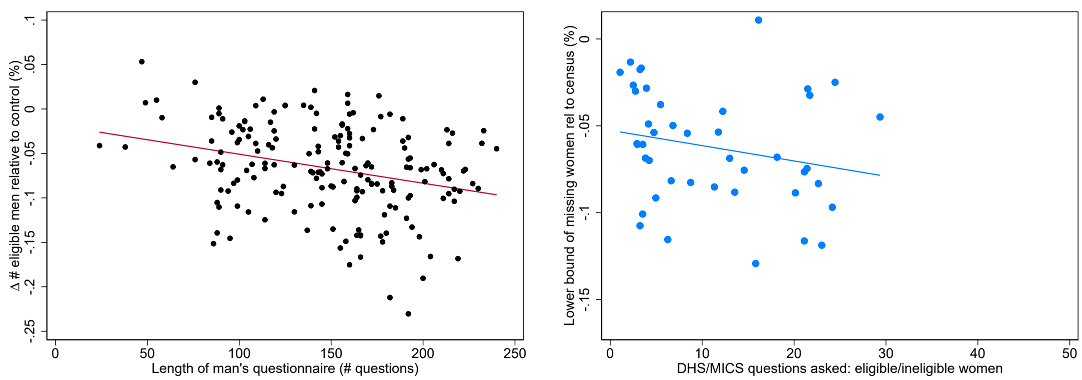
In this section, we present evidence of the robustness of our main finding, the negative effect of the man's questionnaire on the number of eligible men recorded in the household roster. We show that this finding is robust to weighting households in each survey by their respective survey sampling weights, as provided by the DHS and MICS program (see Figure A4a). We also show that surveys which bundle the men's questionnaire with the collection of male biomarkers tend to have more missing



(a) With sampling weights (b) With vs without biomarker collection

Figure A4: Robustness and heterogeneity of effect of man's questionnaire

The figure displays estimates of β from equation (3) relative to the control mean where the outcome is the number of eligible men in the household. In panel (a), households are weighted by their respective sampling weights in each survey. In panel (b), all estimates are unweighted, but surveys that include biomarker collection from eligible men are shown in blue. Circles indicate point estimates; bars indicate 95% confidence intervals. Surveys are sorted in ascending order of the point estimate. Labels as in Figure 3.



(a) Men (b) Women

Figure A5: Effect of questionnaire vs relative question load across surveys

This figure plots estimates of β from equation (3) relative to the control mean against the length of the man's questionnaire across surveys in panel (a), and estimates of the lower bound of the effect of the woman's questionnaire relative to the control mean against the relative question load faced by eligible women relative to ineligible women across surveys in panel (b). See appendix A.2.3 for details on the measurement of the length of questionnaires. The solid lines present a linear fit.

eligible men, in line with increased effort cost of enumerating men (see Figure A4b). Finally, we highlight how more men (and women) are missing in surveys with longer questionnaires, as measured by the number of questions (see Figure A5).

A.3.2 Age displacement, omission and household response

In this section, we examine the relative importance of age displacement of household members relative to their outright omission from household as drivers of missing eligible men in households assigned to the man's questionnaire. Below, we test for differential household response as a third potential margin of manipulation.

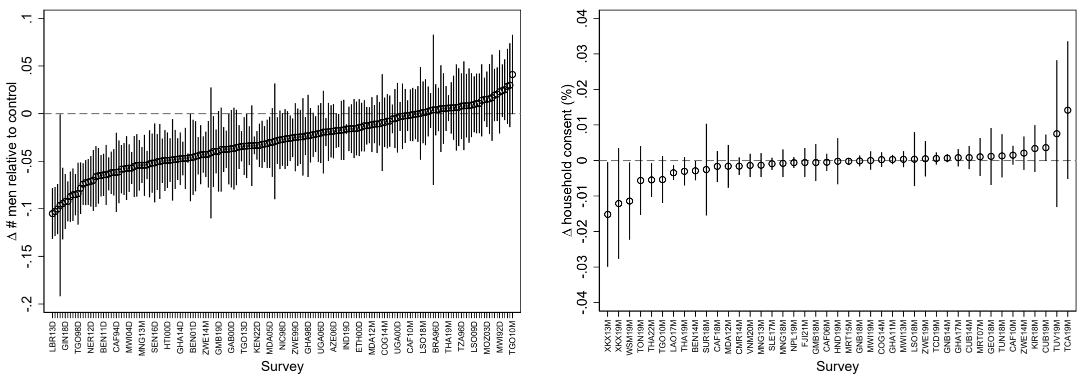
First we show that the total number of men (eligible plus ineligible men) is weakly smaller in households assigned to the man's questionnaire in all surveys (Figure A6a). Second, we present results from pooled regressions showing average effects of the man's questionnaire on eligible and ineligible men and the implied relative contributions of age displacement and the omission of household members from rosters (Table A4). We compute the relative contribution of age displacement to missing eligible as the ratio of the regression coefficients from equation (3) using the number of eligible and ineligible men as respective outcomes:

$$\text{Share explained by age displacement} = \frac{\beta_{ineligible}}{-\beta_{eligible}}$$

The share unexplained by age displacement is attributed to omission. Note that we consider all men outside the eligible age range who are at least 10 years old as ineligible. We set this lower age cutoff to avoid a conflation of our estimates with age displacement from ages below 5 to ages above 5 due to increased question load for young children. However, we vary the lower age threshold to define men of ineligible age to confirm the robustness of our results (compare columns 3, 5 and 6). We also vary whether or not we give equal weight to each survey in our pooled sample by weighting observations with the inverse sample size of their survey (weighted is *Yes*) or not (weighted is *No*). Across all specifications, we find that age displacement can account for approximately 30% of missing eligible men.[57]

Third, we assess the effect of the man's questionnaire on household response. To this end we regress household response on assignment to the man's questionnaire and stratum fixed effects following our main specification. This exercise is limited to surveys from the MICS because the DHS program does not publish any data on households that did not complete the survey. Hence, non-response is not observed for the DHS. In the MICS sample, we find that response is balanced between treatment and control in all but 5 out of 43 surveys (Figure A6b). In all of these five cases, treatment is associated with marginally lower response rates, with the shortfall ranging between 0.3 and 1.4 percentage points. Hence, strategic manipulation of household responses does not appear to be an important margin of data collector response.

[57]Survey-by-survey estimates of the effect of the man's questionnaire on eligible men, ineligible men and total men are provided in Table **??**.



(a) Total number of men in household (b) Household response in the MICS

Figure A6: Effect of man's questionnaire on omission and household response

This figure displays estimates of β from equation (3) relative to the control mean where the outcome is the total number of men in the household in panel (a) and household response across MICS with a randomly assigned man's questionnaire in panel (b). Circles indicate point estimates; bars indicate 95% confidence intervals. Surveys are sorted in ascending order of the point estimate. Labels as in Figure 3.

Table A4: Age displacement vs omission from the household roster

	Eligible men		Ineligible men			
	(1)	(2)	(3)	(4)	(5)	(6)
MQ	-0.074***	-0.077***	0.021***	0.024***	0.023***	0.019***
	(0.0012)	(0.0011)	(0.0010)	(0.0009)	(0.0011)	(0.0008)
Weighted	No	Yes	No	Yes	No	No
Age cutoff	NA	NA	10	10	8	12
Displacement share	NA	NA	0.291***	0.314***	0.307***	0.262***
			(0.0117)	(0.0173)	(0.0160)	(0.0081)
R^2	0.138	0.153	0.099	0.115	0.107	0.097
No. surveys	181	181	181	181	181	181
No. households	3,397,630	3,397,630	3,397,630	3,397,630	3,397,630	3,397,630

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Standard errors in parentheses. All specifications include survey cluster fixed effects. Columns (2) and (4) are weighted by inverse survey sample size.

A.3.3 Selection of data collectors

The eligibility of a given household for the man's questionnaire is revealed on the first page of the household questionnaire. In response to this information, supervisors can strategically assign data collectors to households with and without a man's questionnaire. This raises the question how the eligibility of a household for the man's questionnaire affects the identity of the data collector recording the household roster. Leveraging information on the characteristics of data collectors from the DHS fieldworker questionnaire, available for 19 surveys in our sample, we empirically test how data collector characteristics differ between households with and without a man's questionnaire.[58] In results available from the authors, we find that in most surveys, data collectors in charge of the household roster are significantly less likely to be female in treatment households. The tendency to assign male data collectors to households with a man's questionnaire can be attributed to the survey program's objective to conduct same-sex individual interviews, i.e., to have male (female) data collectors administer man's (woman's) questionnaires. This implies that a male data collector is required at households that are eligible for the man's questionnaire, but not at ineligible households. The effect of the man's questionnaire on age and education varies across surveys, both in sign and magnitude. Experience with previous DHS is negatively affected in most surveys, but also strongly positively in a few surveys.

In the face of these changes in data collector characteristics, it is important to note that, consistent with the idea of moral hazard, selection of data collectors cannot explain the reductions in the number of eligible men as point estimates are barely affected by the inclusion of data collector fixed effects (see Figure A7).

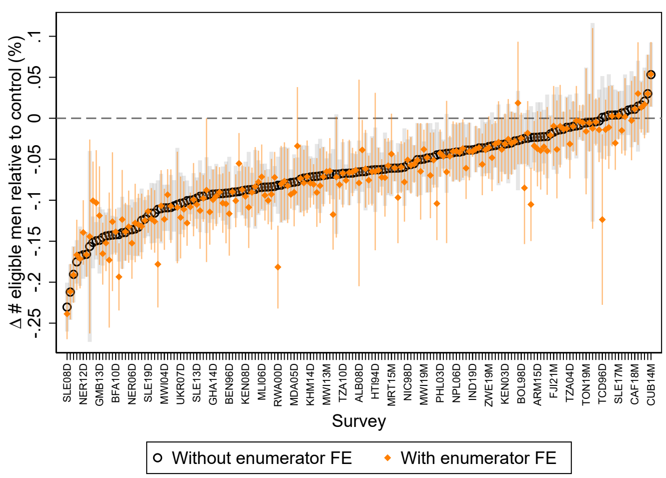


Figure A7: Within-data collector effect of man's questionnaire on # of eligible men

This figure displays estimates of β from equation (3) relative to the control mean where the outcome is the number of eligible men in the household. Circles indicate point estimates and bars indicate 95% confidence intervals. Surveys are sorted in ascending order of the point estimate excluding data collector fixed effects.

A.3.4 Selection of respondents

The assignment to the man's questionnaire may alter the identity of the respondent to the household roster. In fact, in results available from the authors, we find that in almost all surveys, the respondent in man's questionnaire households is less likely to be female, more likely the household head, somewhat older and more educated.[59]

[58]The DHS fieldworker questionnaire was introduced in 2015. Hence, data collector information is not available for earlier surveys. The MICS does not publish any data collector characteristics.

[59]In principle, the change in respondent could affect the listing of household members as respondent selection can affect survey responses (Dillon & Mensah, 2024; Masselus & Fiala, 2024).

A.4 Missing women: additional results

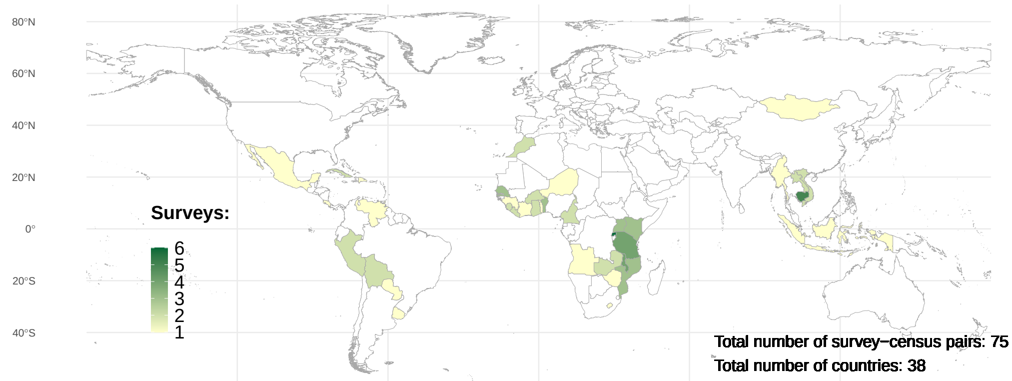
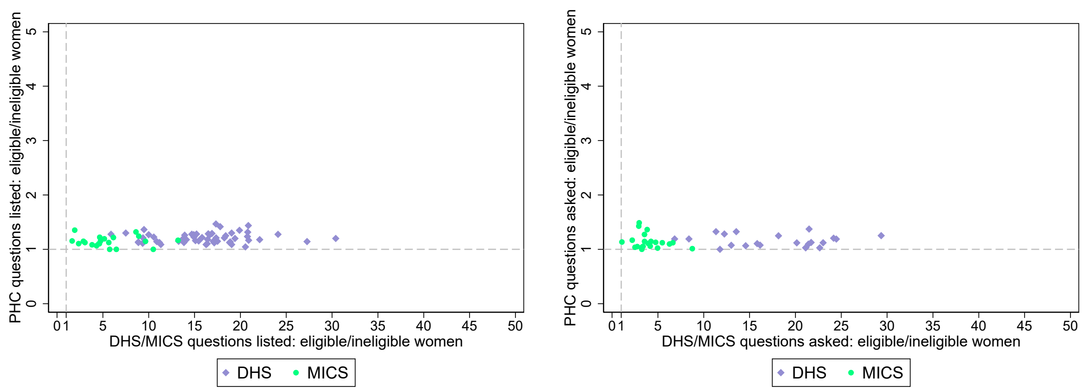


Figure A8: Geographic coverage of DHS/MICS-census pairs



(a) Total number of questions listed (b) Mean number of questions asked

Figure A9: Question load of eligible women relative to ineligible women

This figure plots the question load of eligible women relative to ineligible women in the DHS/MICS against the same ratio in matched population and housing censuses (PHC). In Panel (a), question load is measured by the total number of questions listed in the in the roster and the woman's questionnaire. In Panel (b), it is measured by the mean number of question answered about women of eligible and ineligible age. Panel (a) includes data on all 21 MICS-census pairs and all 46 DHS-census pairs. Panel (b) excludes 23 DHS-census pairs. See Appendix A.2.3 for more information.

A.4.1 Lower and upper bound of missing women

We use the following regression specification to estimate the difference-in-differences of interest:

$$Y_{is} = \beta_0 + \beta_1 SVY_i + \beta_2 Eligible_s + \beta_3 (SVY_i \times Eligible_s) + \mu_{is} \tag{6}$$

where Y_{is} is the number of women of eligibility status $s \in \{eligible, ineligible\}$ recorded in household i. Women are considered eligible if they are in the age range that is eligible for the DHS/MICS woman's questionnaire (usually 15 to 49). They are considered ineligible if they are outside this age range and at least 10 years old. The

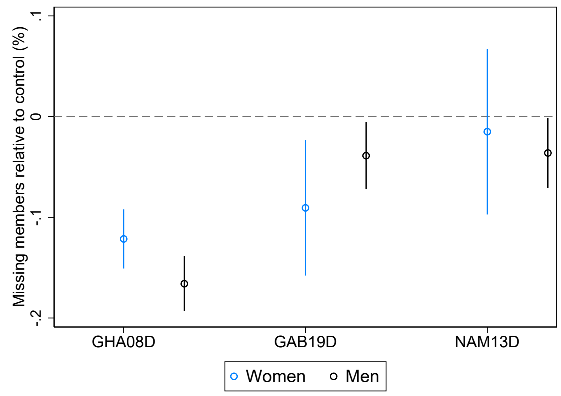
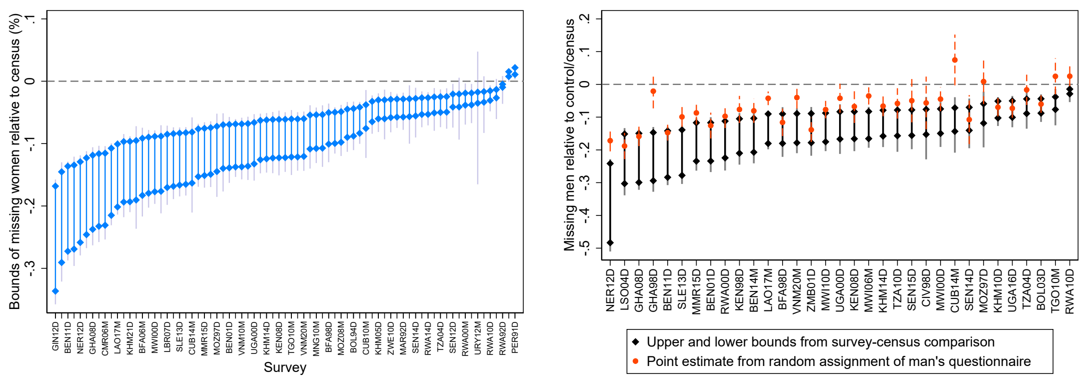


Figure A10: Effect of woman's/man's questionnaire on number of eligible women/men

This figure displays estimates of coefficients from the regression of the eligible number of women (in blue) and men (in black) on the eligibility of their household for the respective individual (woman's or man's) questionnaire. Circles indicate point estimates; bars indicate 95% confidence intervals. Surveys are sorted in ascending order of the point estimate on the number of eligible women. Labels as in Figure 3.



(a) Bounds of missing women

(b) Bounds of missing men (vs estimates)

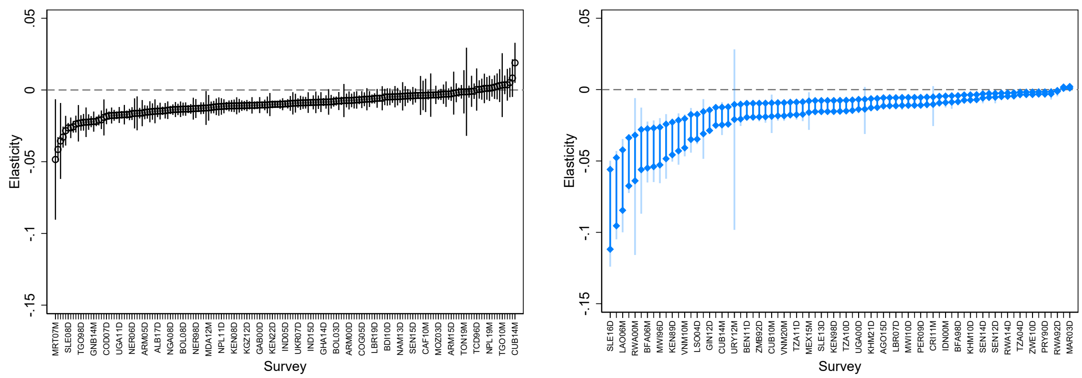
Figure A11: Bounds of missing women and missing men (vs point estimates)

This figure displays estimates of the upper and lower bounds of missing women, indicated by blue diamonds in panel (a), and missing men, indicated by black diamonds in panel (b), derived from survey-census comparisons described in Section 4.2. Panel (b) also displays point estimates exploiting random assignment of the man's questionnaire, indicated by orange circles, in the subset of main sample surveys for which a census for comparison exists. Grey (dashed orange) bars indicate 95% confidence intervals of the bounds (estimates).

lower bound of 10 limits the conflation of the impact of the woman's questionnaire with the impact of the high question load for children under 5 on ineligible women. SVY_i is an indicator that takes the value one if the household roster was recorded by the DHS/MICS and zero if it was recorded by the census. $Eligible_s$ is an indicator that takes value one if the outcome is the number of eligible household members, and zero if it is the number of ineligible household members. We scale survey sampling weights such that the total number of households in surveys and contemporaneous censuses is identical, and cluster standard errors at the household level. β_3 captures the difference-in-differences of interest, the upper bound of missing women. The lower bound is equal to $\beta_3/2$. Figure A11 plots the resulting bounds for women (and men).

A.4.2 Elasticity of sample size with respect to question load

We present the elasticity of sample size with respect to question load for each survey in Figure A12. The elasticities are derived as detailed in Section 4.2. In Figure A13 we benchmark upper and lower bounds of elasticities for women with point estimates for men in the subset of surveys where both can be derived.



(a) Elasticity of sampled men (b) Elasticity of sampled women

Figure A12: Elasticity of sampled individuals with respect to question load

This figure displays estimates of the elasticity of sampled men with respect to question load in panel (a), and sampled women with respect to question load in panel (b). Point estimates for men are indicated by black circles, upper and lower bounds of the elasticity for women are indicated by blue diamonds, with the area between the bounds also colored in blue. Black (grey) bars indicate 95% confidence intervals for men (women).

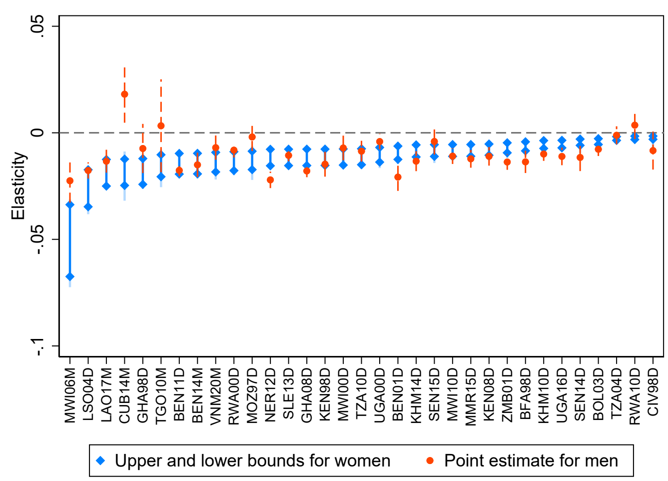


Figure A13: Elasticity of sampled individuals w.r.t. question load: women vs men

This figure displays estimates of the elasticity of sample size with respect to question load. Upper and lower bounds of the elasticity of sampled eligible women with respect to the question load of women are indicated by blue diamonds. The area between the bounds is also colored in blue. Grey shaded bars indicate the 95% confidence intervals of the bounds. Point estimates of the elasticity of sampled eligible men with respect to the question load of men are indicated by orange circles. Dashed orange bars indicate the 95% confidence intervals of these estimates. Surveys are sorted in ascending order of the point estimate of the lower bound. Labels as in Figure 3.

A.5 Mechanisms

A.5.1 Effort cost

In this section, we present additional evidence of the interaction between the man's questionnaire and temperature as well as humidity. While the Princeton Meteorological Forcing Dataset allows us to compute wet bulb temperature in the same way as closely related papers, such as Geruso and Spears (2018) and LoPalo (2023), and thus facilitates comparisons, it does not provide any data after 2010. To be able to include all geo-identified surveys in our sample in the analysis, we instead combine temperature data from CHIRTS-ERA5[60] with rainfall data from CHIRPS v3.[61] The former dataset provides daily wet bulb globe temperature (WBGT) across the globe at the 0.05 degree grid cell level, the latter daily rainfall estimates across the globe at the 0.25 degree grid cell level. We assign WBGT and rainfall to each cluster-day in our sample based on cluster GPS coordinates and interview dates, yielding a combined dataset of nearly 2.4m households across 104,000 survey clusters from 99 surveys.

It is important to note that wet bulb *globe* temperature (WBGT) differs from wet bulb temperature although both measures indicate heat stress. Wet bulb temperature is solely based on dry bulb temperature and humidity, and captures how effectively the human body can cool itself via sweating. WBGT is a more comprehensive heat stress index that additionally factors in solar radiation. While both are measured in degrees Celsius, their scales differ: 30°C wet bulb temperature is near dangerous survival limits, whereas 30°C WBGT is less severe – still indicating high heat stress.

In contrast to results reported in Section 4.3, we use WBGT in this supplementary analysis. This is due to data availability: daily global WBGT is directly available from the Climate Hazards Center, whereas daily global wet bulb temperature is not.[62]

Using regression specification (4), we uncover a analogous relationship between temperature and the effect of the man's questionnaire on eligible men as reported in the main body. As Figure A14 shows we estimate approximately 0.082 missing men in response to the man's questionnaire at WBGTs between 18°C and 22°C, considered most comfortable. The effect increases to 0.086 between 22°C and 26°C and reaches 0.100 as temperatures climb beyond 26°C. For temperatures below 18°C, estimates are noisier and we cannot reject similar effect sizes as observed between 18°C-22°C.

A.5.2 Detection probability

In this section, we show that proxies for a lower probability of detection of household roster manipulation are associated with less missing eligible men in response to the man's questionnaire.

[60] CHIRTS-ERA5 Repository https://doi.org/10.15780/G2F08J (2025), accessed: March 27, 2026.

[61] Climate Hazards Center Infrared Precipitation with Stations version 3 (CHIRPS3) Data Repository: https://doi.org/10.15780/G2JQ0P (2025), accessed: March 27, 2026.

[62] Note that both measures have been used in the economics literature. See LoPalo (2023) for an example of wet bulb temperature and Adhvaryu et al. (2020) for an example of WBGT.

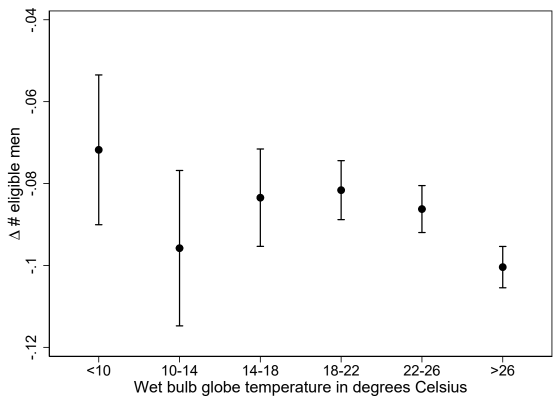


Figure A14: Missing men by wet bulb globe temperature bin

This figure displays the estimates of the regression coefficients β_j from specification 4 across temperature bins. Bars indicate 95% confidence intervals. Standard errors are clustered at the survey cluster level.

First, we leverage cross-survey variation in the use of mandatory re-interviewing (also referred to as audits), field check tables (FCT) and tablets instead of paper questionnaires to examine their association with the effect of the man's questionnaires. We focus on these three survey features because they are systematically documented in survey reports. We manually code them up for all 181 surveys in our main sample, as detailed in Section A.2.5. As Table A5, column 5 shows, surveys without any of the three survey features are missing 8 percentage points more eligible men in response to the men's questionnaire. In contrast, the use of mandatory re-interviewing is associated with a large reduction in missing eligible men by about 5 percentage points. The use of tablets rather than paper forms is only associated with a small reduction in missing men by less than 1 percentage point and field check tables are negatively correlated with missing men, conditional on the other survey characteristics.

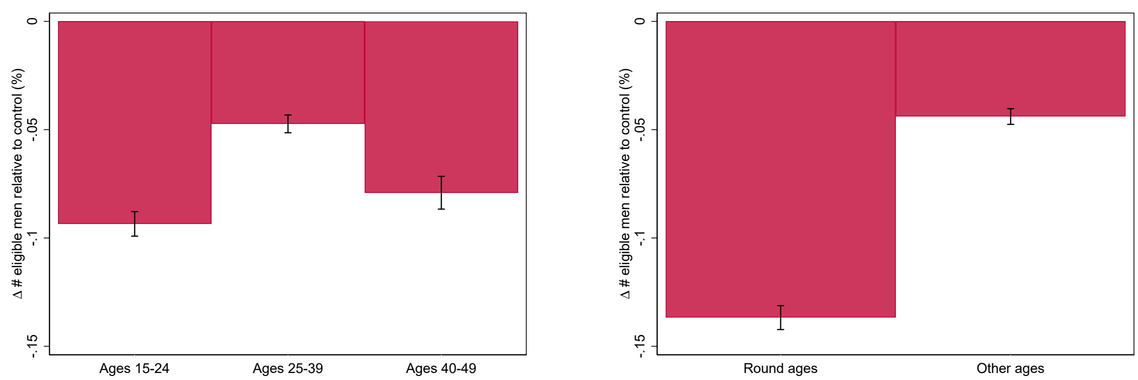
Table A5: Quality control measures and missing men

	Number of eligible men				
	(1)	(2)	(3)	(4)	(5)
MQ	-0.074***	-0.080***	-0.079***	-0.083***	-0.080***
	(0.0012)	(0.0013)	(0.0019)	(0.0015)	(0.0020)
Audit X MQ		0.054***			0.049***
		(0.0034)			(0.0042)
FCT X MQ			0.007***		-0.006**
			(0.0025)		(0.0027)
Tablet X MQ				0.025***	0.009***
				(0.0024)	(0.0031)
R^2	0.138	0.138	0.136	0.138	0.137
No. surveys	181	181	178	181	178
No. households	3,397,630	3,397,630	3,340,797	3,397,630	3,340,797

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. All specifications include survey cluster fixed effects. Standard errors are clustered at the household level.

Second, we examine heterogeneity of the effect of the man's questionnaire by the recorded age of eligible men. We find that effect sizes are larger for age groups for which age manipulation is plausibly harder to detect for supervisors and easier to negotiate with survey respondents. In particular, we detect larger effects among age groups close to the lower and upper eligibility thresholds, for whom age manipulation is arguably easier because it requires a smaller deviation from the truth. In Figure A15a, we estimate that 9.3% and 7.9% of eligible men within 10 years of the lower and upper age threshold, respectively, are missing as a result of the man's questionnaire. In contrast, only 4.7% of eligible men are missing for the age group in between.

We provide additional evidence along similar lines by comparing effect sizes for men of round and non-round ages. This is motivated by a long-standing literature showing significant age heaping at round ages (ending in 0 or 5) in many low- and middle-income countries likely because respondents to household rosters do not know the exact ages of all household members. We hypothesise that manipulating the age of household members whose exact age is not known by respondents is easier. Consistent with this, we find that the effect of the man's questionnaire is much larger among men of round ages (13.7%) than non-round ages (4.4%), as illustrated in Figure A15b.



(a) By age group

(b) Round vs non-round ages

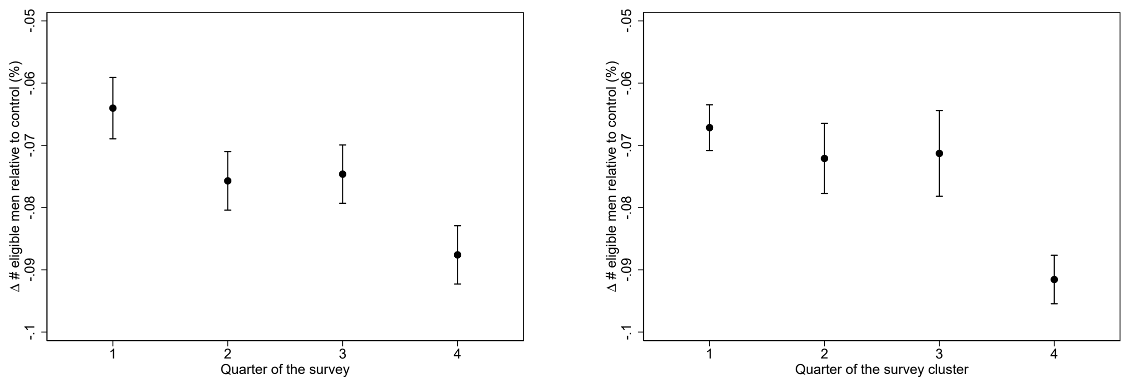
Figure A15: Missing men by age

This figure displays estimates of coefficients from pooled regressions of the number of eligible men of specific age groups on assignment to the man's questionnaire. In Panel (a), we compare estimates for men within 10 years of the lower age eligibility threshold (labelled 15-24), men within 10 years of the upper age eligibility threshold (labelled 40-49 although the upper threshold is higher than 49 in some cases) and men of ages in-between these two age groups (labelled 25-39 although this includes higher ages as well if the upper age eligibility threshold is larger than 49). In Panel (b), we compare estimates for men of round ages (ending in 0 or 5) to estimates for men of non-round ages. Bars indicate 95% confidence intervals. All specifications include survey cluster fixed effects.

A.5.3 Other testable implications

In this section, we provide three additional pieces of evidence consistent with predictions from our theoretical framework.

First, we show that endogenous sample selection becomes more pronounced as surveys progress, consistent with a decreasing continuation value of data collector contracts. To this end we group the households visited by each data collector throughout the entire duration of survey fieldwork into four equal-sized bins by interview time. Hence, the first bin contains the 25% of households visited first, and so forth. Pooling the data from all surveys, we find that more men are missing in response to the man's questionnaire as the survey progresses. As shown in Figure A16a, the share of missing men is 6.4% in the first quarter of all households visited by enumerators and increases to 8.8% in the last quarter.



(a) By overall survey progress

(b) By progress within survey cluster

Figure A16: Missing men by survey progress

This figure displays the estimates of the coefficients from pooled regressions of the number of eligible men in the household on assignment to the man's questionnaire by survey quarter in panel (a), and by data collection progress within survey cluster in panel (b). Bars indicate 95% confidence intervals.

Second, we show that endogenous sample selection becomes more pronounced as data collection progresses within survey clusters, consistent with the importance placed on adherence to pre-determined travel schedules of data collection teams between clusters. To this end, we compare the effect of the man's questionnaire in households interviewed early and late within survey clusters, splitting interviews into four equal-sized bins within clusters by data collector. We estimate that the share of missing men in the first three quarters is approximately 7%, but jumps to more than 9% in the final quarter of households within cluster when data collection has to be wrapped up for travel to the next survey cluster (see Figure A16b).

Third, we examine heterogeneous effects of the man's questionnaire by household size. Since we do not observe the true size of households assigned to the man's questionnaire, we use the presence of polygamous marriage in the household as a proxy for large household size. Focusing on 15 West African countries where polygamy is common, we identify polygamous households as those where at least one woman indicates a positive number of co-wives in the individual interview of the DHS. We verify that polygamous households are indeed larger than monogamous households in our sample: the average polygamous household in the control group has 9.1 members while the average household in the treatment group has only 5.8 members. In results available from the authors, we find that the share of missing men in polygamous households is significantly larger (8%) than in monogamous households (11%). This finding is again consistent with time pressure faced by data collectors due to emphasis on adherence to survey schedule as important driver of screening out individuals.

A.6 Selection: additional results

This section provides additional results on the selection of screened out household members who are eligible for individual questionnaires.

To assess the selection of eligible men, we run the following OLS regression:

$$MQ_{ic} = \alpha_c + \beta X_{jic} + \epsilon_{ic} \tag{7}$$

where MQ_{ic} is an indicator variable that takes value one if household i is eligible for the man's questionnaire, zero otherwise. X_{jic} is a characteristic of eligible member j of household i reported in the household roster and α_c is a set of stratum fixed effects.

Table A6 shows the results from separate pooled regressions of assignment to the man's questionnaire on six different characteristics. These are the estimates illustrated in Figure 7. Note that the underlying sample of surveys differs across outcomes because only few characteristics (relationship to household head and education) are collected in the household rosters of all surveys in our sample. Some characteristics are only collected in a small set of surveys, such as chronic illness or individual-level poverty. Moreover, questions about these can be constrained to specific age subgroups, e.g., poverty is only elicited for 15-17 year olds (see Appendix A.2.4).

In Table A7, we report estimates from joint regressions of assignment to the man's questionnaire on the different individual characteristics to show that our findings are robust to the inclusion of other characteristics. Note that some of the characteristics we study are only available in mutually exclusive subsets of the surveys. Therefore, we report results from a variety of specifications including different combinations of regressors on different subsets of the data. Overall, we find that most coefficients are similar across specifications and comparable to those reported in Table A6.

To understand selection among women eligible for the woman's questionnaire, we compare average characteristics of eligible women in the DHS/MICS to those in the census. To this end, we harmonise age, relationship to household head, years of schooling and marital status between DHS/MICS and censuses (see Section A.2.4).

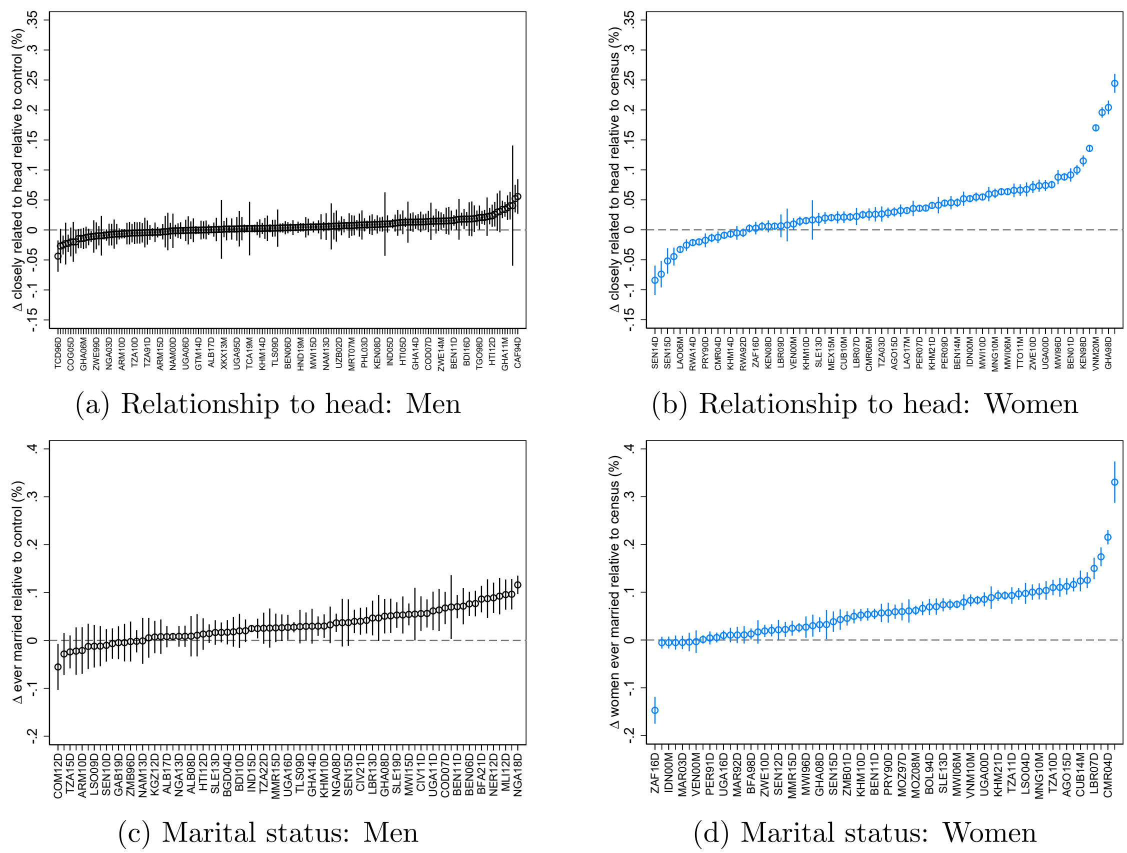
We find very similar patterns of selection for men and women: screened out individuals tend to be more distantly related to the household head and less likely to have ever been married (see Figure A17 for survey-by-survey estimates).

Finally, in Figures A18a and A18c, we plot these estimates of selection for men against our estimates of missing eligible men to confirm that selection is more pronounced in surveys with more missing individuals. We find very similar patterns for

Table A6: Selection on observables: separate regressions

	Dependent variable: Assigned to man's questionnaire					
	(1) Distantly related	(2) No schooling	(3) Never married	(4) Disabled	(5) Chronically ill	(6) Poor
Independent variable	-0.009***	-0.009***	-0.010***	-0.004	-0.015**	-0.052***
	(0.001)	(0.001)	(0.000)	(0.007)	(0.007)	(0.013)
Relative effect	-0.030***	-0.030***	-0.035***	-0.010	-0.035**	-0.127***
	(0.003)	(0.002)	(0.002)	(0.018)	(0.018)	(0.031)
R^2	0.256	0.257	0.323	0.068	0.053	0.235
No. surveys	181	181	77	8	9	5
Age restriction	None	None	None	None	18-59	15-17
No. obs (N)	3,769,808	3,749,485	2,666,040	96,333	105,537	11,539

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$. All specifications include survey cluster fixed effects. Standard errors are clustered at the household level. The independent variable in each specification is specified in the column header.



(a) Relationship to head: Men

(b) Relationship to head: Women

(c) Marital status: Men

(d) Marital status: Women

Figure A17: Selection on observables

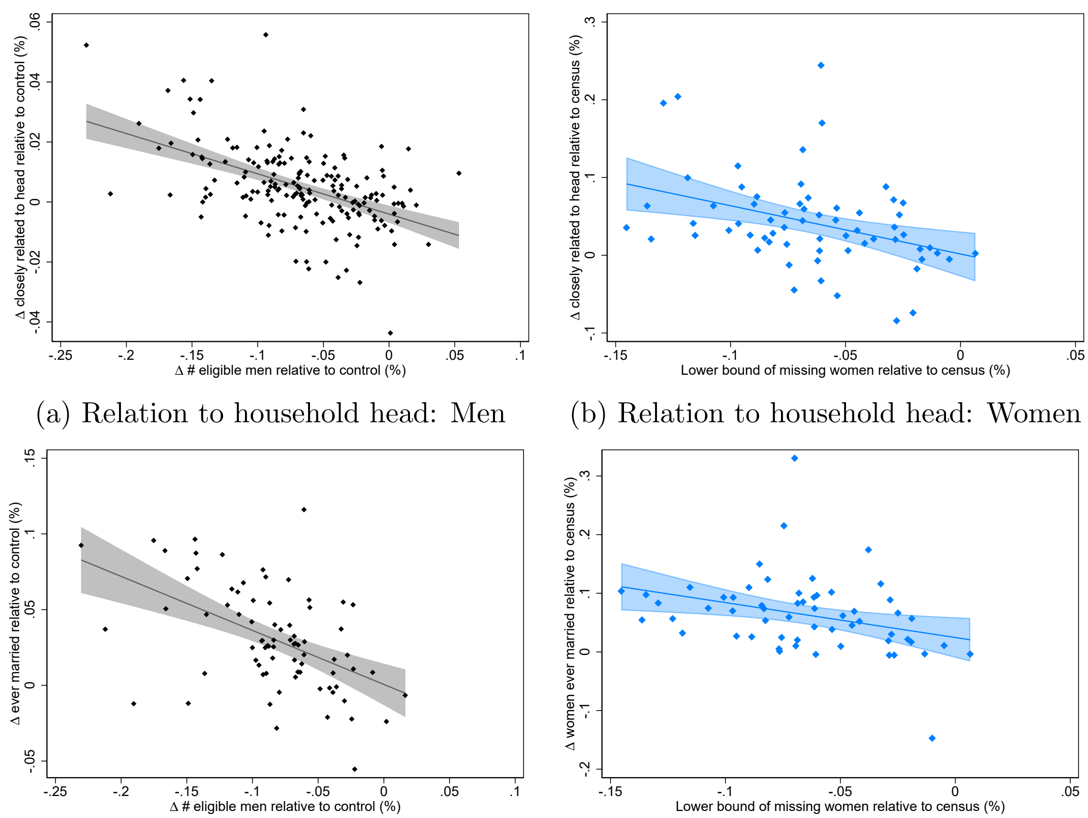
This figure displays estimates of the effect of household assignment to the man's (left) and woman's questionnaire (right) on the characteristics of eligible men and women relative to the relevant comparison group. Standard errors are clustered at the household-level. Circles indicate point estimates; bars indicate 95% confidence intervals.

women in Figures A18b and A18d: limited evidence for selection in cases with few missing women, and stronger evidence of selection in cases with more missing women both in terms of relationship to household head and marital status.

Table A7: Selection on observables: joint regressions

	Dependent variable: Assigned to man's questionnaire							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Distantly related	-0.009***	-0.006***	-0.021***	-0.014**	0.009	-0.015**	-0.013	0.012
	(0.001)	(0.001)	(0.006)	(0.006)	(0.014)	(0.007)	(0.008)	(0.015)
No schooling	-0.009***	-0.010***	-0.019***	-0.028***	-0.015	-0.023***	-0.036***	-0.016
	(0.001)	(0.001)	(0.006)	(0.005)	(0.024)	(0.006)	(0.006)	(0.026)
Never married		-0.011***				-0.018***	-0.021***	-0.131**
		(0.000)				(0.004)	(0.005)	(0.053)
Disabled			-0.005			-0.008		
			(0.007)			(0.007)		
Chronically ill				-0.013*			-0.014	
				(0.007)			(0.012)	
Poor					-0.051***			-0.053***
					(0.013)			(0.014)
R^2	0.257	0.324	0.069	0.053	0.234	0.062	0.036	0.243
No. surveys	181	77	8	9	5	7	5	5
Age restriction	None	None	None	18-59	15-17	None	18-59	15-17
No. obs (N)	3,749,175	2,655,164	95,667	104,575	11,519	82,994	63,573	11,005

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$. All specifications include survey cluster fixed effects. Standard errors are clustered at the household level.



(a) Relation to household head: Men

(b) Relation to household head: Women

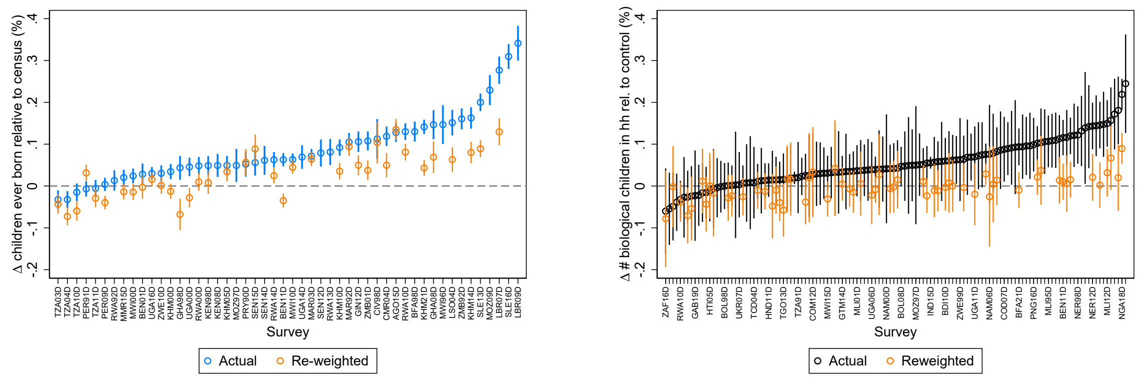
(c) Marital status: Men

(d) Marital status: Women

Figure A18: Selection on observables vs missing individuals

This figure plots in panels (a) and (c) [(b) and (d)] coefficients from regressions of individual characteristics of eligible men on household assignment to the man's questionnaire [differences between survey and census] on the y-axis. Coefficients from regressions of the number of eligible men on household assignment to the man's questionnaire [lower bound of missing women] are plotted on the x-axis in panels (a) and (c) [(b) and (d)]. Each dot represents a survey. The grey (blue) line represents a linear fit for men (women), the shaded area the 95% confidence interval.

A.7 Implications for policy



(a) Fertility of women (census comparison)

(b) Fertility of men (DHS RCT)

Figure A19: Bias in aggregate fertility statistics and selection correction

This figure displays estimates of the effect of household assignment to the man's (right, black) and woman's questionnaire (left, blue) on measures of fertility of eligible men and women relative to the relevant comparison group, i.e., control households for men and census households for women. See Appendix A.2.4 for details on the outcome variables. Regression results using reweighted data are shown in orange. Standard errors are clustered at the household-level. Circles indicate point estimates; bars indicate 95% confidence intervals. Surveys are sorted in ascending order of the respective point estimate. Labels as in Figure 3.

A.8 Implications for research

A.8.1 Climate shocks

We follow Burke et al. (2015) and Corno et al. (2020) to construct a measure of droughts. First, we assign all geo-referenced DHS clusters in our data to 0.5×0.5 arc degree grid cells.[63] Using gridded rainfall data from the Climate Hazards group Infrared Precipitation with Stations (CHIRPS) version 2.0 dataset, we construct annual rainfall for each grid cell-year in our data. Then, we define a drought as calendar year rainfall below the 15th percentile of a grid cell's long-run rainfall distribution. Using the entire CHIRPS time series from 1981 until 2024, we fit a gamma distribution of calendar year rainfall for each grid cell. Then, we use the estimated gamma distribution for a given cell to assign each calendar year's rainfall realisation to its corresponding percentile in the distribution.

Heat is measured by the number of days with a maximum temperature above 30°C in the calendar year of the survey following Somanathan et al. (2021). We construct this measure by spatially matching DHS clusters to grid cells in the CPC Global Unified Temperature dataset from the year of the survey.[64]

[63] At the equator, this corresponds to an area of approximately 2,500 square kilometers.

[64] This dataset is provided by the NOAA PSL, Boulder, Colorado, USA, and was retrieved from their website at https://psl.noaa.gov on Feb 26, 2026.

A.8.2 Institutions and historical legacies

A.8.3 Economic origins

A.9 External validity across survey instruments

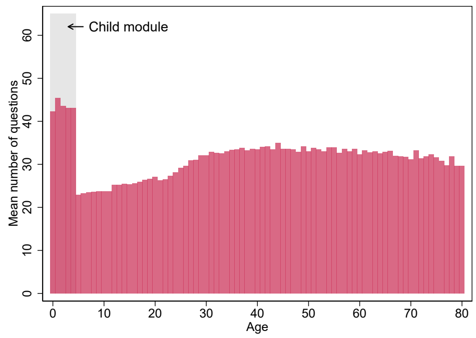
This appendix provides additional graphical evidence and background information on the external validity of our findings to different data-collection instruments.

A.9.1 Endogenous sample selection in other household surveys

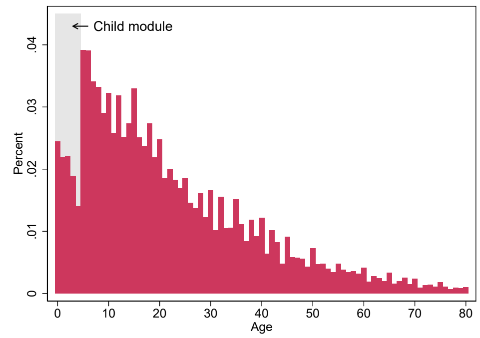
We provide five examples of graphical evidence in line with endogenous sample selection in different types of household surveys. Figure A20 provides four examples from low- and middle-income countries and Figure A21 provides suggestive evidence for a longitudinal survey from a high-income country, the 1997 US National Longitudinal Survey of Youth (NLSY97).

The NLSY97 is heavily used in economics, particularly in the field of labour, featuring in at least 13 top general interest economics papers published since 2010. The process to identify and recruit NLSY97 respondents, however, created an incentive for data collectors to screen out eligible youth. Screening interviews for the NLSY97 were conducted in 1997 and 1998. The same interviews were also leveraged to screen for eligible participants for the Profile of American Youth (PAY97). All youth aged 12 to 16 were eligible to participate in the NLSY97, while those aged 18 to 23 were eligible for the PAY97. Similar to other household surveys, the NLFS97 employed a 'screen-and-go' procedure that created an incentive for data collectors *not* to identify eligible youth because these were a lot more work for them.

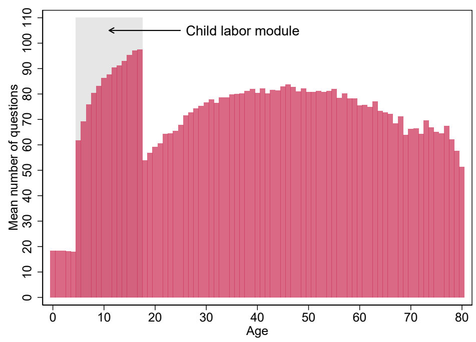
Examinations of the screening results reveal a pronounced under-coverage of youth of eligible ages (Horrigan et al., 1999; Moore et al., 2000). As Figure A21a shows, screened households have a lot fewer youth aged 12 to 23 than comparable households in the March 1997 CPS, strikingly similar to Figure 1 and consistent with endogenous sample selection. Moore et al. (2000) also provide evidence of selection on observables: family income of youth in the NLSY97 is lower than that of households with resident youth aged 12 to 16 in the CPS (see Figure A21b).



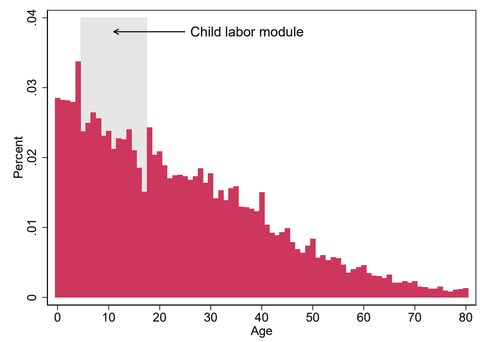
(a) Zambia LCMS 2015: Question load



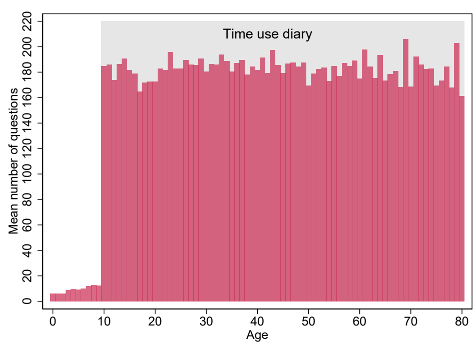
(b) Zambia LCMS 2015: Age distribution



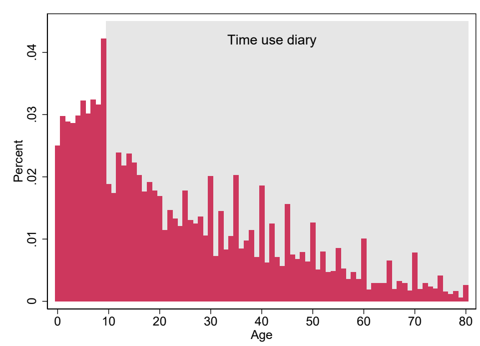
(c) Tanzania ILFS 2014: Question load



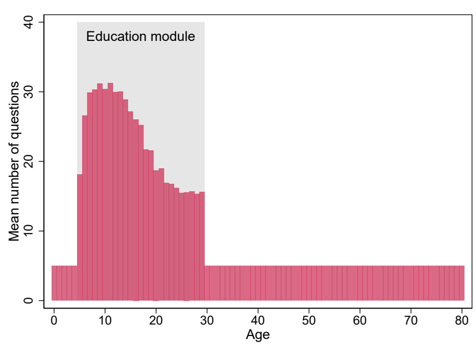
(d) Tanzania ILFS 2014: Age distribution



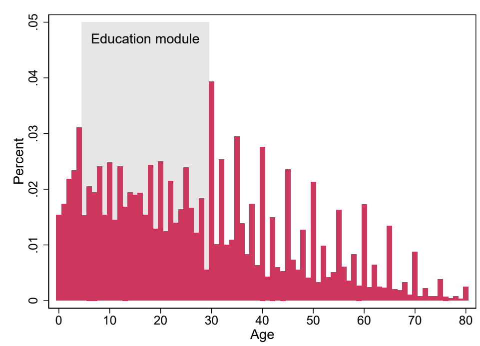
(e) Ghana TUS 2009: Question load



(f) Ghana TUS 2009: Age distribution



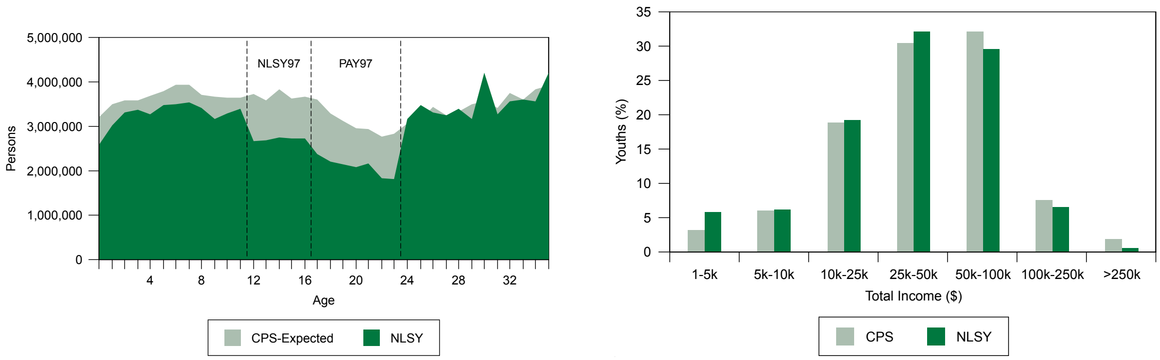
(g) India NSS 2007: Question load



(h) India NSS 2007: Age distribution

Figure A20: Question load and age distribution in education surveys

This figure shows the distribution of the mean number of questions asked about household members by age in the left hand side panels against age distributions in the right hand side panels. Panels (a) and (b): 2015 Zambian Living Conditions Monitoring Survey (LCMS); panels (c) and (d): 2014 Tanzanian Integrated Labour Force Survey (ILFS); panels (e) and (f): 2009 Ghana Time Use Survey (TUS); panels (g) and (h): 2007 Indian National Sample Survey (NSS, 64th round). Shaded areas indicate survey modules that are only applied to specific age groups.



(a) Age distribution in screened households

(b) Distribution of youth by family income

Figure A21: Sample selection in the National Longitudinal Survey of Youth 1997

The left panel compares the age distribution of household members in the households screened for the NLSY97 and the PAY97 to the age distribution in the March 1997 CPS. The right panel compares the family income of youth in the NLSY97 to the family income of households with resident youth aged 12 to 16 in the March 1997 CPS. Both figures adapted from Moore et al. (2000). Panel (a) corresponds to Figure 5.5, panel (b) corresponds to Figure G-1.

A.9.2 Endogenous sample selection in firm surveys

Firm and farm size distributions are commonly used to study labour and capital market frictions and misallocation. However, many of the underlying firm/farm censuses use size thresholds to determine sample inclusion. In particular, many censuses limit the amount of information collected about small units, incentivising data collectors to manipulate unit sizes to fall below size thresholds or to omit larger units entirely.

For example, the design of the Indian Economic Census' (EC) creates an incentive for data collectors to manipulate firm size. Aiming to record all formal and informal non-farm businesses in the country, data collectors visit all buildings in the entire country, recording any firm and their basic characteristics, including the number of employees. Additional information is then collected for firms above a given size.

We exploit the shifting of the eligibility threshold for additional data collection between consecutive censuses to reveal firm size manipulation by data collectors. In fact, in 1998 no such threshold existed – the amount of information collected about firms was independent of their size. In the 2005 Economic Census, the requirement to complete an address slip for all firms employing 10 or more workers was introduced (Ministry of Statistics and Programme Implementation, India, 2005). In the next firm census in 2013, the eligibility threshold was adjusted downward to a firm size of 8 and the additional information requirement was expanded to include a description of a firm's major activity and its source of registration alongside their name and address. All of this information was collected on a form labelled 'Schedule 6C'.

Figure 10 illustrates bunching of firms below the respective eligibility thresholds of 10 and 8 in 2005 and 2013, and does not reveal any sign of bunching in the 1998 firm size distribution. In fact, in results available from the authors, we show that the firm size distribution in India closely follows a power law, a regularity previously observed in many other countries (Axtell, 2001; Hernández-Pérez et al., 2006). In contrast, the 2005 and 2013 distributions clearly display excess mass to the left of their respective threshold and missing mass to the right of it. Moreover, bunching moves in accordance with the change in threshold between these two years. We interpret this as evidence of firm size manipulation by data collectors. To put the scale of missing employees from under-reported firm sizes into context, we fit a linear model to describe the relationship between the log share of establishments and the log number of employees. We estimate that the observed manipulation reduced recorded employment by approx. 6.3 and 2.8 million workers in 2005 and 2013, respectively.[65]

Our findings have implications for at least two separate streams of literature. First, our findings are relevant for structural work on labour and capital market frictions, which frequently uses moments of the firm size distribution for calibration or structural estimation. The firm size distribution in the Indian EC, for example, is leveraged to study labour regulation (Amirapu & Gechter, 2020), microfinance (Buera et al., 2020) and female entrepreneurship (Chiplunkar & Goldberg, 2024), which we have shown to be susceptible to distortions introduced by data collectors.

Second, the uncovered manipulation of firm size by data collectors has implications for reduced form work evaluating effects on non-farm employment, as recorded in the EC. This includes recent work on the impact of roads (Asher & Novosad, 2020), electricity (Burlig & Preonas, 2024), canals (Blakeslee et al., 2023; Asher et al., 2024), public employment programs (Muralidharan et al., 2023) and politics (Asher & Novosad, 2017). If the treatment is uncorrelated with the share of non-farm workers that are missing due to firm size manipulation, treatment effects will be attenuated because recorded non-farm employment is less sensitive to treatment than actual non-farm employment.[66] If treatment is instead correlated with the share of non-farm workers that are missing due to firm size manipulation, treatment effects can be upward or downward biased depending on whether treatment makes firms more or less susceptible to downward size manipulations by data collectors. For example, treatment could generate growth among initially small firms, thereby shifting them above the size threshold. Data collectors, however, would adjust their recorded employment downwards, thus (partially) masking the growth effects of treatment.

[65] We assume manipulation windows to range from 7 (5) to 30 (28) in 2005 (2013). Our estimates increase if we allow the manipulation window to be larger. We exclude firms with less than 2 and more than 90 workers, the latter due to differential regulation for firms of size 100.

[66] Consider the evaluation of a randomly assigned treatment T on non-farm employment E. Let s be the share of missing workers. The estimated treatment effect β is the difference in mean recorded employment $\bar{E_R} = \bar{E} * (1-s)$ between treatment and control: $\beta = \bar{E_R^T} - \bar{E_R^C} = (1-s)(\bar{E^T} - \bar{E^C})$.

A.10 Discussion

A.10.1 Information-bias trade-off

Table A8: Man's questionnaire and missing men over time

	Length of man's questionnaire		Elasticity of sampled men		Share of missing men	
	(1)	(2)	(3)	(4)	(5)	(6)
2000s	63.9864***	57.9442***	-0.0022**	-0.0018*	0.0196***	0.0181***
	(6.4742)	(6.9010)	(0.0009)	(0.0009)	(0.0045)	(0.0058)
2010s	75.7705***	69.0714***	-0.0014*	-0.0019**	0.0211***	0.0219***
	(5.6666)	(6.9697)	(0.0009)	(0.0009)	(0.0048)	(0.0060)
2020s	102.4518***	103.1996***	-0.0007	-0.0022*	0.0219***	0.0278***
	(10.3637)	(8.9780)	(0.0016)	(0.0013)	(0.0074)	(0.0084)
Country FE	No	Yes	No	Yes	No	Yes
Mean 1990s	103.0357	103.0357	-0.0097	-0.0097	0.0612	0.0612
N	181	181	181	181	181	181
R^2	0.5152	0.7293	0.0265	0.4051	0.1558	0.5122

All specifications include survey program fixed effects. The omitted decade is the 1990s. The length of the man's questionnaire is measured by the number of questions listed in the questionnaire. See Section 4.1 for details on the estimation of missing men and Section 8 for details on the estimation of the elasticity of sampled men. Robust standard errors in columns (1) and (2). Standard errors in columns (3)-(6) are bootstrapped using 100 repetitions.

A.10.2 Comparison with non-response bias

In this section, we examine the relative contribution of endogenous sample selection and non-response to missing data from eligible men in the DHS. To this end, we compare our estimates of endogenous sample selection to non-response rates from eligible recorded men in households with a man's questionnaire. Note that for this exercise, we set missing eligible men due to endogenous sample selection to zero for all surveys for which our point estimate of missingness is negative.

Figure A22 shows the total share of eligible men with missing data in each survey and decomposes it into the contribution from endogenous sample selection and non-response conditional on having been recorded as eligible. On average, both margins contribute to missing data in roughly equal shares.

Interestingly, endogenous sample selection and non-response are negatively correlated across surveys (results available from the authors). This suggests that data collectors screen out some of those men who would not have completed the questionnaire if they had been recorded.

What is the aggregate bias from endogenous sample selection and non-response? To assess this, we compare the characteristics of three groups of eligible men: those listed in households without a man's questionnaire, those listed in households with a man's questionnaire and those in households with a man's questionnaire that completed the man's questionnaire. Focusing on four observable characteristics — age, relationship to the household head, years of schooling and marital status — we find that the biases from both sources of missing men frequently go in the same direction. More specifically, non-response reinforces the selection on each of the four observables

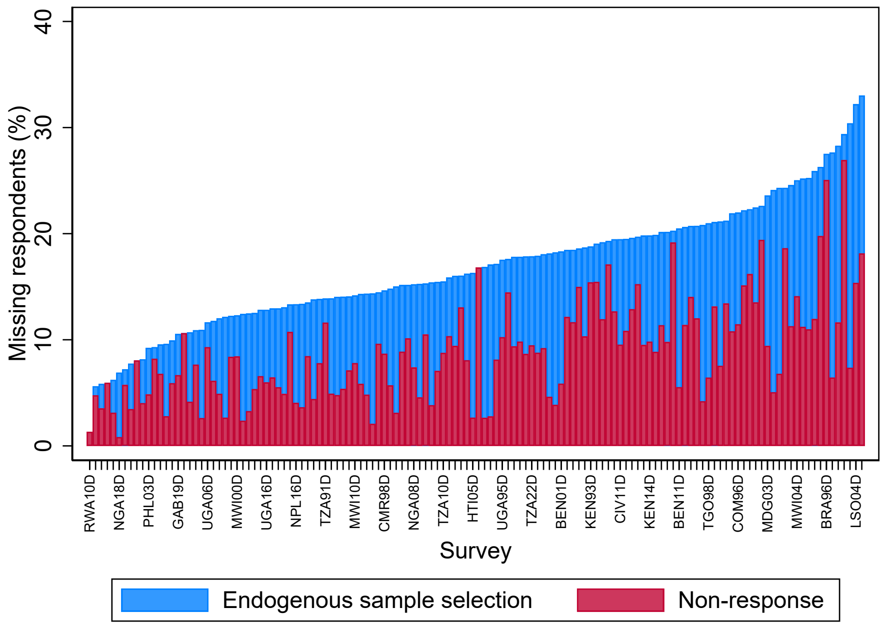
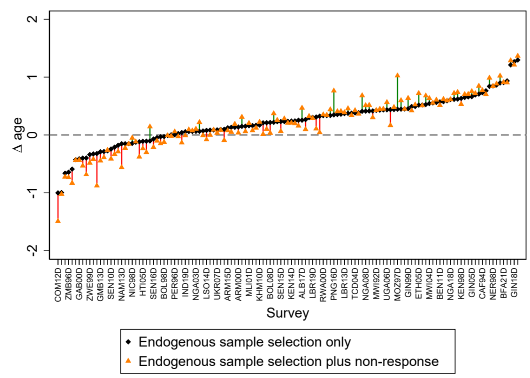


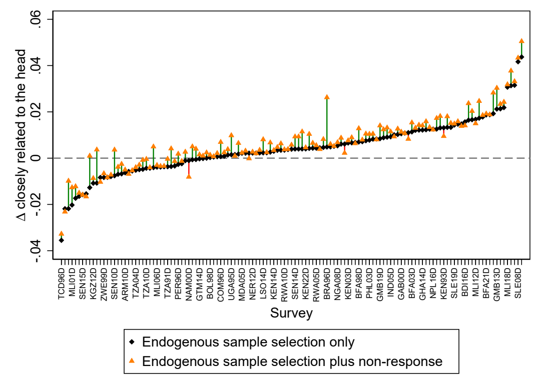
Figure A22: Missing data due to endogenous sample selection and non-response

For each survey, this figure displays the total share of eligible men for whom a completed man's questionnaire is missing, decomposed into the share due to endogenous sample selection (blue) and the share due to non-response (red). Surveys are sorted by the total share of missing. Labels as in Figure 3. Three surveys (BGD 2004, HND 2011, NPL 2011) are excluded because we are not able to replicate the response rate given in the final survey report using the microdata.

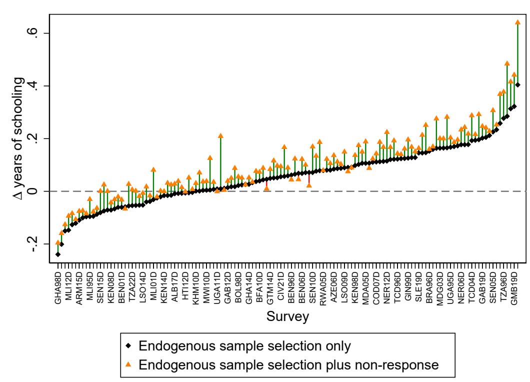
from endogenous sample selection in about two thirds of the surveys in our sample. However, in the remaining third of surveys the bias from both two sources go in opposite directions and cancel each other out, as evident in Figure A23.



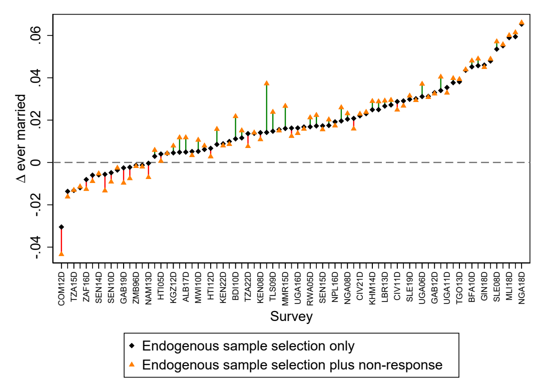
(a) Age



(b) Relationship to household head



(c) Education



(d) Marital status

Figure A23: Aggregate bias from endogenous sample selection and non-response

This figure displays estimates of assignment to the man's questionnaire on four observable characteristics of recorded eligible men in black for all 135 DHS in our sample. Estimates shown in yellow additionally account for non-response by comparing eligible men in the control group to the subset of eligible men in the treatment group that completed the man's questionnaire. Estimates are connected by green lines if the latter estimate is larger than the former and by red lines if the opposite is the case. Surveys are sorted by the estimates of the impact of endogenous sample selection only. Labels as in Figure 3.