UAB

Universitat Autònoma de Barcelona

A Tale of Two Sources: Pooling Census and Survey Microdata for Demographic Research

Kristen Jeffers

Doctoral Thesis submitted for the degree of

Doctor of Demography

A Tale of Two Sources: Pooling Census and Survey Microdata for Demographic Research

Author

Kristen Jeffers

Directors

Dr. Albert Esteve Palós – Dr. Ewa Batyra

Tutor

Dr. Albert Esteve Palós

UNIVERSITAT AUTONOMA DE BARCELONA
DEPARTAMENT DE GEOGRAFIA
CENTRE D'ESTUDIS DEMOGRÀFICS

February 2026

Universitat Autònoma de Barcelona

CED Centre d'Estudis Demogràfics

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisors, Dr. Albert Esteve and Dr. Ewa Batrya. It is not an exaggeration to say this thesis would not exist if it were not for their guidance and patience. When Albert first agreed to supervise my PhD at the Centre for Demographic Studies (CED), I doubt he anticipated he was making a nearly decade-long commitment. Over the course of seven years, Albert provided the clear direction and gentle affirmation I needed to keep going. My advice to his future students is to record your meetings–you do not want to miss any of his wisdom, which often pours out at such a pace that would challenge even the most diligent notetaker. Ewa brought new energy and thoughtful perspectives when they were most needed. Her insights significantly enriched this dissertation.

I would like to thank Dr. Lara Cleveland and Dr. Matt Sobek from IPUMS, who taught me everything I know about microdata and supported my decision to pursue a PhD, while I was supposed to be working for them. I would like to acknowledge Dr.Miriam King who first sparked an interest in demography – only partly thanks to the homemade baked goods she brought to class every week. She was the person to introduce me to IPUMS and she paved the way for my eventual employment there. I am indebted to Lara, Matt and Miriam for their guidance early in my career and for the opportunities they opened up for me.

I am grateful to several people at CED who offered feedback, support, and friendship during my PhD program: Dr. Joan García, Dr. Antonio López, Dr. Elisenda Rentería, Dr. Iñaki Permanyer, Anna Turu, Soco Sancho and Inés Brancos.

Everything I have accomplished is thanks to the support of my parents, who always prioritized education and showed me what it meant to work hard at something worthwhile. Thank you, Dad, for asking about my dissertation frequently enough to show interest but infrequently enough to avoid causing stress. My mom–the literal and figurative cheerleader–had an unconditional belief in me, no matter the endeavour. I wish more than anything she were here for this milestone.

My brother Peter set the academic bar very high from a young age. I have always been trying to keep up with his brilliance.

Thank you to my in-laws for your subtle but persistent encouragement.

To my son Darragh, my favorite demographic outcome, thank you for providing the best reason to finish. I would so much rather spend my weekends with you than in front of the computer.

Finally, to my husband, Tom. It is impossible to articulate the countless ways you support me. Your confidence in me is my greatest motivator. Thank you for patiently entertaining all my threats to quit but making sure I never did.

TABLE OF CONTENTS

A TALE OF TWO SOURCES: INTRODUCTION

During the last 30 years, the volume of data available for sociodemographic inquiry has increased considerably. While demography has always relied on large-scale population data in a way that is unique compared to other social sciences (Kashyap 2021), the format and accessibility of these data have evolved over the last several decades. Driven by advances in computing and widespread digitalization, aggregate and tabular data have given way to anonymized individual- and household-level microdata as the cornerstone of demographic research. The advent of the IPUMS microdata harmonization projects (Ruggles et al. 1995) and the dissemination of standardized microdata files by several international survey programs beginning in the early 2000s fundamentally transformed research design and output in the field of demography and across the social sciences. While individual-level analyses were not uncommon between 1960 and 2000—Kashyap (2021) argues that the paradigm shift from aggregate to individual-level data occurred with the launch of the World Fertility Survey in the 1970s—the digital dissemination of census and survey microdata beginning at the start of the 21st century represented a significant turning point in the quantity and accessibility of data for multivariate demographic research. Indeed, Ruggles (2014) describes the IPUMS-fueled "explosion" in the availability of individual-level population data during the 2000s and 2010s, and Kashyap (2021) suggests that demography experienced a data revolution long before the modern era of "big data".

The data revolution in demography has been characterized by an increase in comparative research. Thanks in large part to the continued expansion by IPUMS of available harmonized census and survey microdata data, researchers can now study large swaths of the globe with relative ease. Scholars increasingly design large-scale analyses to explore trends over time and across countries and regions. Such studies yield critical evidence on how economic, social, political, and environmental conditions shape demographic outcomes by enabling systematic comparison across diverse contexts. To maximize geographic and temporal scope, this type of comparative research sometimes pools data from multiple infrastructures, capitalizing on the flexibility of microdata to construct comparable measures across available census and survey samples (see for example Ackah et al. 2025; Batyra and Pesando 2021; Esteve et al. 2017; Esteve et al. 2020; Esteve and Reher 2021; Esteve and Reher 2024).

At the same time, the demographic data landscape continus to evolve. In wealthy regions, methodological developments are driving a move away from complete, field-based census enumerations to combined methods or fully register-based systems. In low-and-middle-income countries (LMICs), resource and capacity constraints affect traditional data collection efforts. Recent cuts in foreign aid and contributions to international organizations has created uncertainty around the future of international survey programs in LMICs. A potential outcome of these shifts is reduced availability of representative, population-level data from traditional sources. In this context, researchers will continue to make creative use of existing data, pooling data from different instruments and infrastructure to fill gaps and to calibrate new data sources.

It is in this context that my thesis seeks to understand how available harmonized population microdata shape our understanding of recent demographic change. In three distinct essays, I explore–explicitly and implicitly–applications of pooled census and survey data. The first two essays directly assess the complementarity of census microdata and data from international survey programs in LMICs. Less methodological than the first two, the third essay leverages integrated census and American Community Survey data to document the evolution of a rare phenomenon over time. In parallel, I advance knowledge in three key areas of social demography: female educational attainment, household size and composition, and living arrangements among young adults. The three studies are unified by their reliance on harmonized census and survey data from IPUMS.

A TRIBUTE TO IPUMS

The role of IPUMS in democratizing access to census and survey microdata and simplifying comparative analysis cannot be overstated.

The release of the first anonymized census and Current Population Survey microdata samples by the United States Census Bureau in the 1960s and 1970s represented a significant turning point for demographic and social science research (Magnuson and Ruggles 2025). It also helped to institutionalize public use microdata as a standard component of the U.S. national statistical system (Ruggles 2014). These microdata samples quickly became the foundation for demographic research in the United States (Ruggles 2014), though a lack of consistency across U.S. samples in coding schemes and data structure limited their use for comparative work (Magnuson and Ruggles 2022).

When it comes to broader comparative research, the real paradigm shift was the arrival of IPUMS. The original incarnation of IPUMS was launched in 1992 as an integrated database of samples from the 1880 and 1900 to 1990 U.S. censuses (Magnuson and Ruggles 2022). In 1995, the first version of its signature web-based data extract system went online (Magnuson and Ruggles 2022). IPUMS now enabled researchers to build customized datasets that integrated consistently coded samples from multiple years. In the late 1990s, while IPUMS added the Current Population Survey and the American Community Survey to its collection of U.S. microdata, the effort to expand IPUMS to include data from other countries began. Quickly successful, IPUMS International was disseminating anonymized and harmonized census data for more than 70 countries about a decade later (Ruggles 2014). In early 2026, IPUMS International disseminates microdata samples for more than 600 censuses and surveys from 104 countries (IPUMS International 2026b).

Now composed of nine major data collections and 2.6 billion records, IPUMS is the largest database covering the human population (IPUMS 2026). Since 2016 and 2024, respectively, harmonized data for the Demographic and Health Surveys (DHS) and UNICEF Multiple Cluster Indicator Surveys (MICS) are among those available from IPUMS. This thesis leverages this extensive data collection, analyzing data from 192 censuses and 140 surveys from 69 countries.

I had the privilege to spend seven years of my early professional career working as a data analyst on IPUMS International. I can attest firsthand to the rigor of the IPUMS harmonization methodology and the painstaking attention given to ensure that data given a common, harmonized code reflects the same underlying concept across countries. I myself spent countless hours reading instructions to census enumerators and referencing national standards and regulations to determine if brick referred to the same type of building material in Ethiopia as it did in Ecuador, if vocational training captured the same level of education in Portugal as it did in South Africa, that self-employed represented the same status in the labor market in Cambodia as it did in Romania. Thanks to a large and geographically diverse team of research assistants, we had native speakers of the world’s primary languages on hand to check translations and explain nuances of differences between response categories. The researchers downloading pooled, harmonized datasets from the IPUMS website trusted our work, and we took this responsibility very seriously.

The unparalleled value of IPUMS for comparative research derives not only from harmonization, but also from the data integration it facilitates and the constructed variables and documentation it provides. The online data extract system allows researchers to build and download customized, pooled datasets with only the required samples and variables formatted for the statistical software of their choice (Sobek 2016). This thesis exploits variables constructed by IPUMS that indicate family relationships across records that greatly simplify the study of household composition and living arrangements (Sobek et al. 2011). In recent years, IPUMS has also made it easier for researchers to integrate data from multiple infrastructures by providing commonly coded variables and harmonized geographic footprints across some of its data collections including IPUMS International, IPUMS DHS and IPUMS MICS. The extensive, integrated online metadata guide researchers in the design and interpretation of analyses in an integrated and customizable manner. The IPUMS method of data dissemination relieves researchers of countless hours of data cleaning, validation and management.

This thesis is largely motivated by my work with IPUMS. On the one hand, it is a tribute. I am convinced of the almost limitless potential of the harmonized census and survey data available from IPUMS and one underlying aim of this thesis is to showcase their broad and deep applicability. On the other hand, my time with IPUMS underscored the meticulous effort required to make data collected in different places at different times truly interoperable. As a result, this thesis also springs from concern that some researchers may take for granted the comparability of data from different data infrastructures, pooling data for their analyses without anywhere near the same level of scrutiny expected of us in the IPUMS offices in the basement of Wiley Hall.

STRUCTURE AND MAIN THEMES

The three essays that make up my dissertation explore, in one way or another, different drivers and consequences of demographic change: female educational attainment, household size and composition, and living arrangements among young adults. The essays presented in chapters 1 and 2 are primarily concerned with measurement, assessing coherence across international census and survey data. The essay presented in chapter 3 represents an application of integrated U.S. census and American Community Survey data to study a rare demographic phenomenon. The final section provides conclusions.

The first essay, "Trends in Female Education in Low- and Middle-Income Countries: Coherence across Data Sources" examines the coherence of estimates of female educational attainment across three data sources: census samples from IPUMS International, DHS, and MICS. In view of the fundamental role female educational attainment plays in individual and population-level demographic outcomes, as well as its policy relevance for economic and social development, the essay responds to the absence of systematic analyses of comparability across data sources in measures of educational attainment. While the central aims of the essay are to determine how trends vary across data sources and to assess the extent to which these data sources offer a consistent record of progress in education, the study also provides an empirical account of educational expansion among women aged 25–29 across 75 LMICs. The analysis pools almost 20 million individual-level observations from 535 censuses and surveys, covering the period 1960 to 2017. Results contribute to a better understanding of the validity of a critical independent variable in social science research and provide practical information to the research community about the reliability of comparative investigations using three important data sources for demographic and development studies.

The second essay, "Discrepancies in Household Size and Composition Between Census and DHS Data Across 22 Low- and Middle-Income Countries" also examines coherence across data sources. This study compares measures of household size across census samples from IPUMS International and DHS surveys and identifies the origins of inconsistencies. As estimates of household size and structure could indicate systematic differences in the populations being captured across the two data sources, the implications for demographic research are wide-reaching. The study uses both aggregate and individual-level data from 78 censuses and 123 DHS surveys in 22 LMICs collected during the period 1969 to 2019. The analysis documents differences in average household size across data sources and investigates the characteristics of the households and individuals included in the samples that drive discrepancies. The discussion sheds light on differences in the design and implementation of censuses and DHS surveys that may create biases. The results provide important evidence for researchers using these sources to study fertility intentions and outcomes, union formation and living arrangements, migration and urbanization, educational attainment, labor market participation, health, intergenerational relations, gender disparities and socio-economic inequalities.

A departure from the first two essays in several ways, the final essay "Non-family Living Arrangements Among Young Adults in the United States" presents a new geographical focus and substantive rather than methodological research questions. The study documents trends in non-family living arrangements in the U.S. over time and identifies the individual characteristics associated with living in non-family households. The results fill a gap in existing research on changing patterns of family formation and the impact of living arrangements, which focuses primarily on cohabitation among unmarried couples. The study leverages integrated U.S. census and American Community Survey data available from IPUMS USA for the period 1990 to 2019, and represents an application of pooled census and survey data to achieve the longest available view (at the time of writing) of a rare–but increasing–socio-demographic phenomenon. While initially conceived as a cross-national study, it quickly became evident that it is not possible to distinguish non-married partners from other nonrelatives in many international census and survey samples. The essay also showcases the utility of IPUMS-constructed family interrelationship variables which allow the identification of family ties among individuals not related to the head of household. The results provide new knowledge on living arrangements among young adults and open up avenues for future research in this area.

Admittedly, I come to this thesis with a bias towards census data. My professional experiences with IPUMS and the United Nations have convinced me of the centrality of the census for demographic research. With full awareness of limitations that are acknowledged in this thesis, I believe census microdata are one of the most valuable resources available for the study of the complex demographic transformations unfolding around the world. Ostensibly, I take an exploratory approach to questions of data coherence in chapters 1 and 2. In full transparency, these studies are motivated by a slight doubt about surveys as a complement or substitute to the almighty census. Are surveys suitably comparable? Can these sources *really* be used interchangeably? The supremacy of the census is therefore a theme that cuts across this work, and one that I return to in the conclusions of this thesis.

A NOTE ON TIMING, PUBLICATION, AND CO-AUTHORSHIP

I have undertaken my doctoral studies in a *very* part-time manner, writing my dissertation over the course of seven years while working full-time. In the time that has elapsed since each of these essays was written, new, relevant data have become available from DHS, MICS, and IPUMS–a testament to the speed at which these data providers collect and release data.

While this is not officially a dissertation by compendium of publications, two of the three chapters of this thesis have been published in peer-reviewed academic journals. These studies are presented with minimal, mainly stylistic edits to improve consistency across the document. Certain details, therefore, reflect the situation at the time of writing.

Chapter 1, "Trends in Female Education in Low- and Middle-Income Countries: Coherence across Data Sources" was written between 2019 and 2020 and was published in *Comparative Population Studies*, Volume 47 on 8 August 2022. Chapter 3, "Non-family Living Arrangements Among Young Adults in the United States," was written during 2021 and 2022 and published in the *European Journal of Population*, Volume 40 (1) on 6 March 2024. Chapter 2, "Discrepancies in Household Size and Composition Between Census and DHS Data Across 22 Low- and Middle-Income Countries," was written in 2024 and 2025 and will be submitted for publication.

The papers are co-authored with my supervisors Dr. Albert Esteve (essays 1, 2, and 3) and Dr. Ewa Batyra (essays 2 and 3), but I led and carried out the majority of the work. Dr. Esteve and Dr. Batyra provided valuable conceptual and methodological guidance, but the data management, analysis, interpretation of results and writing of the manuscripts reflects my work alone. The term "we" is therefore used throughout the thesis. This approach conforms to the requirements of the School for Doctoral Studies at the Universitat Autònoma de Barcelona (UAB).

CHAPTER 1

TRENDS IN FEMALE EDUCATION IN LOW- AND MIDDLE-INCOME COUNTRIES: COHERENCE ACROSS DATA SOURCES

INTRODUCTION

Education is widely recognized as a vehicle for individual agency, social change and economic growth. It is also considered an important policy instrument for development in many low- and middle-income countries (LMICs), where enrollment in primary and secondary schooling is not universal and gender disparities in education remain. As a result, national and international policy agendas include as key objectives expanded access to schooling and the closing of the gender gaps in education (see United Nations 2013; United Nations Children's Fund 2019; UNESCO 2016; UNESCO 2018). Monitoring progress towards these goals requires comparable data across countries and over time. The availability of international census and survey microdata allows for cross-national comparative assessments of education participation and completion. However, we lack systematic analyses of how trends in education vary across these data sources, and of the extent to which these trends offer a consistent account of progress in education in recent decades. Do the multiple rounds of census and survey microdata available for LMICs provide a coherent picture of trends in educational expansion? Are there systematic differences across data sources? Which specific censuses or surveys stand apart from the trend presented by other census and survey observations in the country?

To answer these questions, we pooled individual-level data from 535 censuses and surveys from 75 low- and middle-income countries to monitor trends in educational attainment and examine coherence across data sources. The data cover 1960 to 2017 and come from three sources: population censuses from IPUMS International, the Demographic and Health Surveys

(DHS), and the Multiple Indicator Cluster Surveys (MICS)[1]. All censuses available from IPUMS International as well as all DHS and MICS surveys collect information on school attendance and educational attainment. We examine trends in educational attainment among females age 25 to 29. While gender gaps in education have reversed in most high and middle income countries, women trail men in school enrollment and completion in many developing countries, particularly in sub-Saharan Africa and South Asia (Ilie and Rose 2016; Kebede, Goujon, and Lutz 2019). Accordingly, international policy agendas include specific targets to address gender disparities in education. We find strong coherence across data sources overall with heterogeneity by data source and region. Our results provide practical information to the research community about the validity of comparative investigations using the data sources we examine. Our analysis focuses on young women, but sensitivity analyses examining trends in male educational attainment yield similar results and conclusions.

BACKGROUND

The contribution of basic education to economic growth is widely recognized (Schultz 1961; Becker 1994; Barro and Lee 1994; Cohen and Soto 2007). Its importance is not, however, confined to economic progress. Education has intrinsic positive effects for people and societies. Education imparts knowledge and skills that provide access to better paying jobs, alleviating poverty and increasing the availability of cultural and economic resources for households (Becker 1994). Highly educated individuals tend to have better health and live longer lives (Hannum and Buchmann 2005; Kc and Lentzner 2010; Lutz et al. 2014; Masquelier and Garbero 2016). The children of highly educated parents are also more likely to have better health and cognitive outcomes (Caldwell 1981; Cochrane, Leslie and O'Hara 1982; Lutz et al. 2014; Bicego and Boerma 1993; Rosenzweig and Wolpin 1994; Mellington and Cameron 1999; Wang 2003; Schady 2011; Abuya, Onsomu, Kimani and Moore 2011; Vikram, Vanneman and Desai 2012; Grépin and Bharadwaj 2015). Due to the link between female education and demographic outomes, educational expansion has the potential to significantly influence population growth (Lutz and Skirbekk 2014). Based on this evidence, policymakers and development practitioners view access to and completion of education as key policy instruments for development in low- and middle-income countries. The United Nations 2030

[1] To access international census data from IPUMS International, visit international.ipums.org. DHS data are available for download from the DHS Program, dhsprogram.com, and IPUMS DHS, idhsdata.org. MICS data are available for download at mics.unicef.org.

Agenda for Sustainable Development includes educational expansion and gender parity in schooling as primary goals, challenging LMICs to ensure universal and equal access at all levels of education. Many of these countries trail high-income countries in quantitative and qualitative progress in education. Enrollment in primary and secondary education is far from universal and college graduates represent less than five percent of the population in much of sub-Saharan Africa (UNESCO Institute for Statistics 2019). Moreover, the expansion of primary, secondary and tertiary education has not been gender neutral (Grant and Behrman 2010, Dorius and Firebaugh 2010, Dorius 2013; UNESCO 2019).

During the Millenium Development Goals (MDGs) era, countries with better performance monitoring made better progress towards targets (Jacob 2017). To track progress towards new education targets, accurate measures of school participation and completion are fundamental. Policy makers and researchers worldwide have traditionally relied on official reports and statistics from international agencies and some scholarly databases (Barro and Lee 2013, for example) to obtain macro level indicators on educational attainment. These statistics were compiled from censuses and surveys, resources for which microdata were rarely available to researchers. Only recently have researchers gained access to census and survey microdata from LMICs upon which to build new and large comparative analysis. These microdata come from three main sources: the IPUMS International database for population censuses and household surveys (Minnesota Population Center 2020), the Demographic Health Surveys Program (ICF 2004-2017), and the Multiple Indicators Cluster Surveys (UNICEF 2020).

Among these sources, censuses have served as the underlying source for global databases on educational attainment like those maintained by the United Nations. They also provide the benchmark observations for modelled estimates and projections of educational attainment produced and used by the scholarly community (Barro and Lee 2013; Bauer et al. 2012; Cohen and Soto 2007; Kc, Barakat, Goujon, Skirbekk, Sanderson, and Lutz 2010; Lutz et al. 2014; De la Fuente and Doménech 2002; Jordá and Alonso 2017). For many countries, especially LMICs, population censuses represent the only available source for educational data. Compared to surveys, censuses offer some advantages—most importantly universal coverage of the population—which contribute to reliable and consistent trends over time. However, in the best of cases, censuses are conducted every ten years. In low-resource and/or conflict settings, more than ten years may elapse between censuses. Furthermore, access to census microdata has been limited until recent years. The IPUMS International project has simplified access to census microdata by collaborating with National Statistical Offices to compile, integrate, and harmonize representative samples of census microdata from around the world. In early 2020, harmonized microdata from more than 300 censuses and 98 countries are available to researchers free of charge through IPUMS International. Nearly every country in South and Central America disseminates microdata through IPUMS International. Twenty-seven of 54 African countries disseminate data through IPUMS International. Twenty-two of 60 low- and middle-come countries in Asia, the Middle East, and the Pacific disseminate through IPUMS.

Prior to the publication of international census microdata by IPUMS, cross-national research on the expansion and implications of education in LMICs mostly relied on surveys. Compared to censuses, surveys provide more timely data and cover a broader range of topics in greater depth. Data from two global survey programs are most frequently used to monitor development progress in LMICs: the Demographic and Health Surveys (DHS) and the Multiple Indicator Cluster Surveys (MICS). The Demographic and Health Surveys are nationally representative household surveys that collect data on a number of topics related to population, health, and nutrition (ICF 2019). DHS surveys are conducted about every 5 years. Since 1984, more than 400 DHS surveys have been completed in over 90 countries in Africa, Latin America, Asia, Oceania, and Eastern Europe. Data on literacy, school attendance, and educational attainment are collected for all household members in every DHS survey. DHS surveys primarily target women in reproductive years and they have been the main source for large cross-national studies in trends in women's education and their influence on family transitions and health (e.g. Castro-Martín 1995, Lloyd 2005, Grant and Behrman 2010).

A complement to DHS, the UNICEF MICS Surveys provide data on the wellbeing of women and children in LMICs. More than 300 surveys have been fielded in 116 countries covering the period 1993 to the present (UNICEF 2020). Most MICS surveys are also nationally representative and collect education information similar to that collected in DHS surveys. DHS and MICS are important sources for measuring progress towards the Sustainable Development Goals (SDGs). Likewise, the data are used extensively by social scientists and policy and health researchers to study a wide range of topics related to child health and wellbeing and human development (e.g. Pace, Lee and Grogan-Kaylor 2019; Kang, Aguayo, Campbell, West Jr, 2018; Jeong, Bhatia and Fink 2018). DHS and MICS surveys use similar multi-stage cluster sampling strategies. Primary sampling units (PSU) typically correspond to census enumeration areas and are stratified by administrative regions and urban and rural areas before selection. A complete household listing is conducted in each selected PSU and households are selected for the survey by equal probability systematic sampling (ICF International 2012). This type of probability sampling relies on updated and reliable sampling frames. Censuses are considered the most suitable sampling frames for DHS and MICS surveys. For this reason, we expect strong coherence between census-based estimates and survey-based estimates. When a census has not been conducted recently, alternative sampling frames such as electoral rosters are used.

Used together, IPUMS, DHS and MICS constitute a vast repository of individual-level data for cross-national research on population dynamics, development, and health in LMICs. Despite the nearly limitless potential of these data, cross-national studies combining census, DHS, and MICS microdata are still rare because microdata for a critical mass of countries only became available in the last decade. This paper—a first attempt in this direction—examines the feasibility of combining these sources to examine trends in and implications of educational expansion. While the topical coverage and policy purposes of these data collection instruments differ, censuses and DHS and MICS surveys share basic features that make them suitable for large comparative studies. The collection of information on basic educational attainment is one of these common features. In theory, comparing educational attainment across countries and data sources should be straightforward. Most education systems are organized around three distinct levels: primary, secondary, and tertiary. Systems within regions and among countries with shared colonial histories are usually quite similar. Moreover, the International Standard Classification of Education (ISCED) provides clear guidelines for measuring educational attainment in censuses, surveys, and population registers. In practice, accurate comparisons across countries are difficult because the definition of educational levels varies significantly across countries.

Scholars considering education at the global level have grappled with comparability across data sources. In their work to produce population projections disaggregated by level of education, researchers at the Wittgenstein Centre for Demography and Global Human Capital have relied on microdata from censuses, DHS, MICS, and other household surveys to construct global datasets on educational attainment (Bauer et al. 2012; Goujon et al. 2016; Lutz et al. 2014; Lutz et al. 2018; Speringer et al. 2019). Bauer et al. (2012) and Speringer and colleagues (2019) document the comparability issues and harmonization challenges encountered in the constructuction of the base-year datasets used across these education and population projections, including discrepancies between national education systems and ISCED categories, changes over time in national education systems, inconsistencies across data sources in the treatment of complete versus incomplete levels, age-heaping in survey data, and a lack of detail at the lowest and highest levels of education in developed and developing countries, respectively. Results of validation exercises confirm that different data sources lead to different educational compositions, and that census or register data are generally the most reliable. The few examples provided of discrepancies between DHS surveys and censuses and intra-country inconsistencies across DHS surveys support the need for further investigation of coherence across across these important sources of population data.

This paper seeks to fill this gap by providing a systematic evaluation of trends in educational attainment across the three primary sources of demographic data in low- and middle-income countries: population censuses, Demographic and Health Surveys (DHS), and UNICEF Multiple Cluster Indicator Surveys (MICS). We aim to identify coherence and inconsistencies in the empirical measurement of education within and across data sources and to contribute to a better understanding of the validity of a critical independent variable in development research.

DATA AND METHODS

We use individual-level data from 210 IPUMS census samples[2], 219 DHS surveys[3], and 106 MICS surveys to assess coherence in the measurement of education across data sources in 75 low- and middle-income countries. For a robust picture of trends over time, we include in the analysis all countries with four or more censuses and/or surveys across two or more data sources. The data were collected during the period 1960 to 2017 and cover more than half of the 138 countries designated as low or middle income by the World Bank. A plurality of countries studied are represented in all three data sources: IPUMS, DHS, and MICS (table 1.1).

IPUMS census samples vary in size but typically include 10 percent of the population. Sample sizes for DHS and MICS surveys typically cover 20,000 to 50,000 households. See table A1.1 in the Appendix for analysis sample sizes. DHS and MICS data used in the analysis were collected in household questionnaires and come from corresponding household-member

[2] Certain microdata samples disseminated by IPUMS come from large-scale household surveys rather than censuses. See Appendix for more information.
[3] Certain DHS samples accessed from IPUMS DHS. See Appendix for more information.

microdata files. DHS and MICS surveys collect basic demographic information on all usual residents and visitors in surveyed households via the household questionnaire. From these census and survey samples, we select women age 25 to 29. In many low- and middle-income countries, women trail men in access to and completion of secondary education. Development agendas focus on confronting barriers to girls' schooling and require accurate information on female educational attainment. Women in this age range represent the population that recently completed formal education, more accurately reflecting the educational context in the country at the time of the survey compared to older cohorts. Coherence analysis based on men yield similar results.

Table 1.1 Number of analysis countries by data source and world region

	Africa	Asia	Europe	Latin America / Caribbean	**Total**
DHS, MICS, IPUMS	17	5	1	3	26
IPUMS, DHS	7	8	–	7	22
IPUMS, MICS	2	4	1	9	16
DHS, MICS	8	2	–	1	11
TOTAL	34	19	2	20	**75**

Source authors’ elaboration

Most IPUMS census samples and DHS and MICS surveys include retrospective questions on years of schooling and highest level of education attended and/or completed. Our analysis focuses on the population completing secondary and higher education. Existing literature suggests secondary education is an important threshold for individual outcomes and development (Caldwell and McDonald 1982; Ainsworth, Beegle and Nyamete 1996; Desai and Alva 1998; Abuya, Ciera and Kimani-Murage 2012; Makoka and Masibo 2015). We generate a dichotomous educational attainment variable that distinguishes women who have completed secondary education or higher from those who have not. For census samples, we recode the IPUMS EDATTAIN variable. DHS survey data were accessed directly from the DHS Program and from IPUMS DHS (Boyle et al. 2020). For DHS surveys, we use the educational attainment summary variable constructed by the DHS Program during data processing (hv109 or EDSUMM in IPUMS). The MICS datasets do not include a summary educational attainment variable. We construct a summary measure of educational attainment for MICS samples according to each country's educational system using variables indicating the highest education level attended and highest grade completed at level.

We calculate the percentage of women age 25 to 29 who have completed secondary education for each census or survey, yielding 535 country-year observations in total. The mean number of census and survey observations among countries examined is 7.7. We use polynomial regression to fit trends over time in educational attainment within each country. More specifically, we have specified a second-degree polynomial (quadratic) regression model to account for the curvilinear nature of the relationship between female educational attainment and time. Previous studies of global educational expansion during the time period of interest have identified a sigmoidal pattern in most countries, with initially slow growth, a phase of rapid expansion, and then a slow approach towards universal participation (Barakat and Durham 2014). We explored both quadratic and cubic specifications, but found the quadratic model to provide the most parsimonious fit to the data for the developing countries included in our analysis, many of wich have not yet reached the final phase of educational expansion.

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 Year^2 + \beta_2 Year + \varepsilon$$

In the model specified above, P is the proportion of women age 25-29 completing secondary or higher education. Each observation of P is weighted by sample size. By doing this, we give more importance to census observation over DHS and MICS. Census microdata samples are usually drawn from complete-count census databases. Results derived from census samples are therefore extremely close to those derived from complete-count databases. We use the parameters estimated by the model to generate predicted values of P. We then compare observed and predicted values, calculating and summarizing absolute differences between the two and classifying observations based on absolute deviance from the predicted trend. We disregarded the use of confidence intervals as samples vary greatly in size and large samples usually yield very narrow confidence intervals such that even a tiny deviation from that value represents an outlier. On the contrary, small sample sizes generate wider confidence intervals and relatively large deviations may fall within the interval.

RESULTS

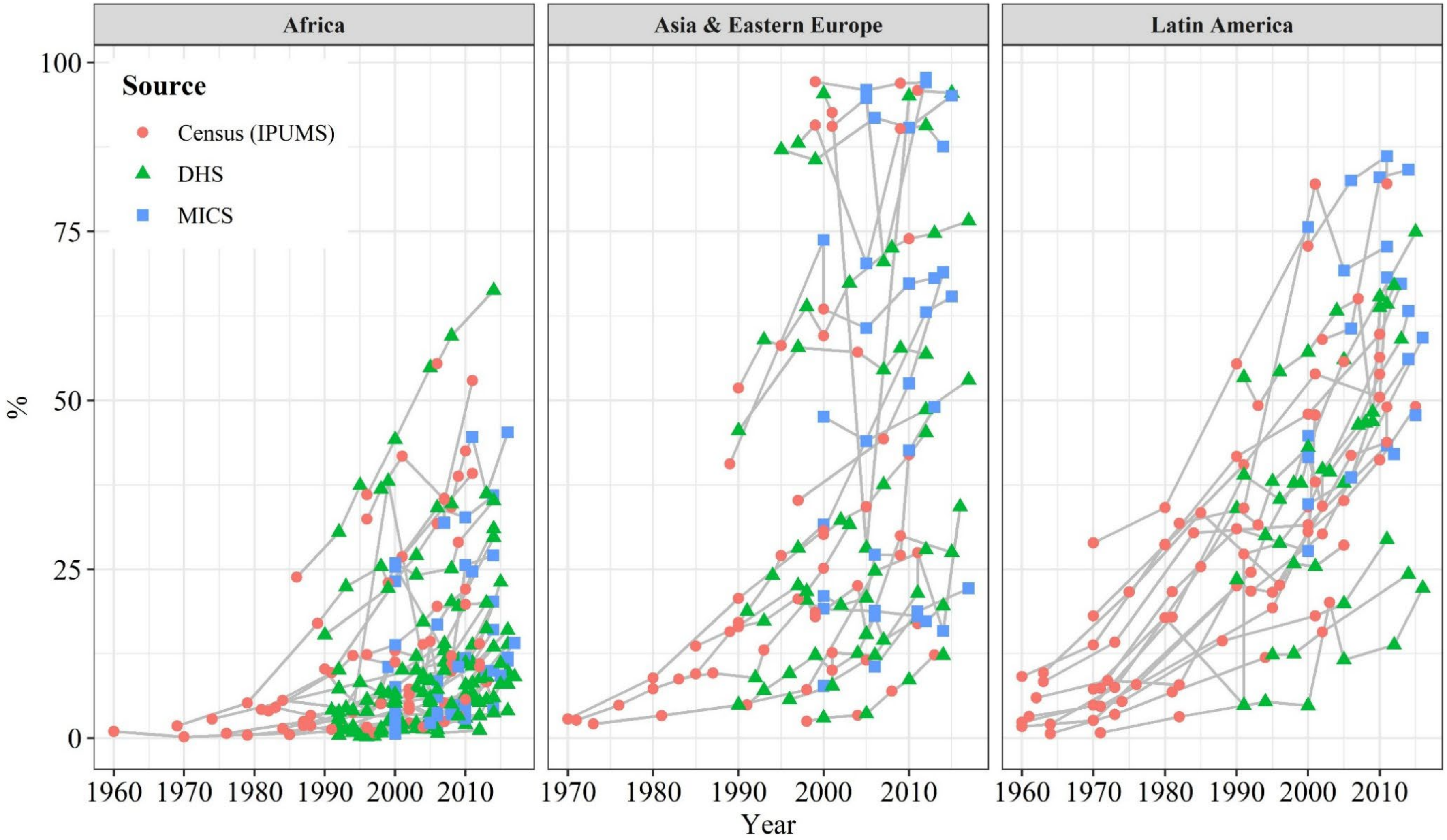
OBSERVED TRENDS IN EDUCATIONAL ATTAINMENT

Figure 1.1 presents by country and world region estimates of the percent of women age 25 to 29 completing upper secondary education or higher for the 75 countries included in our analysis (see also Appendix table A1.1). Each line represents a country, showing change over time based on multiple data sources. The color and shape of the symbol marking each point estimate indicates the data source: Census, DHS or MICS. This figure corroborates trends well known to researchers and policy makers. Over the last several decades, in all world regions, there has been a general upward trend in the proportion of young women completing at least secondary education in LMICs. The percent of women age 25 to 29 completing secondary education or higher has increased between the earliest census or survey observation and the most recent in nearly every country examined.

However, levels of secondary or higher completion vary widely across countries and regions. Educational attainment is lowest in Africa. Less than 10 percent of women age 25 to 29 had completed secondary education according to the most recent census or survey in Burkina Faso, Burundi, Benin, Central African Republic, Chad, Ethiopia, Mali, Mozambique, Niger, Senegal, and Togo. Levels of secondary completion are much higher in Latin America. Since 2010, in two-thirds of the Latin American countries studied more than 50 percent of women complete secondary or higher. Levels of completion were highest in Cuba (2014) and Trinidad and Tobago (2011), where more than 80 percent of young women complete secondary or higher, and lowest in Haiti (2016) and Guatemala (2014), where less than 25 percent of young women complete secondary or higher. Recent levels of secondary completion vary within Asia, where the most heterogeneity is observed. Female educational attainment remains low in several of the Southern and Southeast Asian countries studied according to recent surveys. Less than 30 percent of women age 25 to 29 had completed secondary or higher in Cambodia (2014), Bangladesh (2014), Laos (2017), India (2015), and Pakistan (2012). In contrast, educational attainment is high in Central Asia and Eastern Europe: more than 95 percent of women age 25 to 29 complete secondary education or higher in Armenia, Belarus, Kazakhstan, Kyrgyz Republic, and Ukraine according to recent surveys.

The rate of increase in secondary completion in LMICs during the last several decades also varies across countries and regions, as reflected in the variation in the slope of country trend lines in figure 1.1. Change has been slowest in Africa, where the median annual increase in the percentage of women age 25 to 29 completing secondary or higher among the countries examined is 0.43 percentage points. By comparison, the median annual increase in the percentage of women age 25 to 29 completing secondary or higher is 0.66 percentage points among Asian countries examined is 1.1 percentage points among Latin American countries studied.

Figure 1.1 Percent of women age 25-29 completing secondary education or higher by year, country, data source and world region



Source Authors' elaboration based on data from IPUMS International, IPUMS DHS, DHS Program, and MICS

The dominant trend of increasing female educational attainment over time and the relative levels of educational attainment we observe in recent censuses and surveys reflect what we understand about access to education across and within world regions. A closer look at country-specific patterns, however, reveals inconsistencies within and across data sources, even among observations that are only a few years apart. We do not expect that educational expansion is linear in all countries. This expectation is reflected in the data. In some countries, secondary or higher completion has expanded more rapidly in recent decades and we see this trend reflect in a J-shaped curve for many countries. In other countries, educational attainment increased rapidly during the MDG-era but has slowed in recent years as countries approach universal secondary participation. This trend is reflected in an inverse J-shaped curve. We do, however, expect that in most countries educational attainment among young women will always be non-decreasing. As clearly visible in figure 1.1, this is not the case. In 49 of 75 countries examined, at least one point estimate of secondary completion among women age 25 to 29 is lower than the previous point estimate. In 11 countries, we observe a decline of 10 percentage points or more between adjacent point estimates.

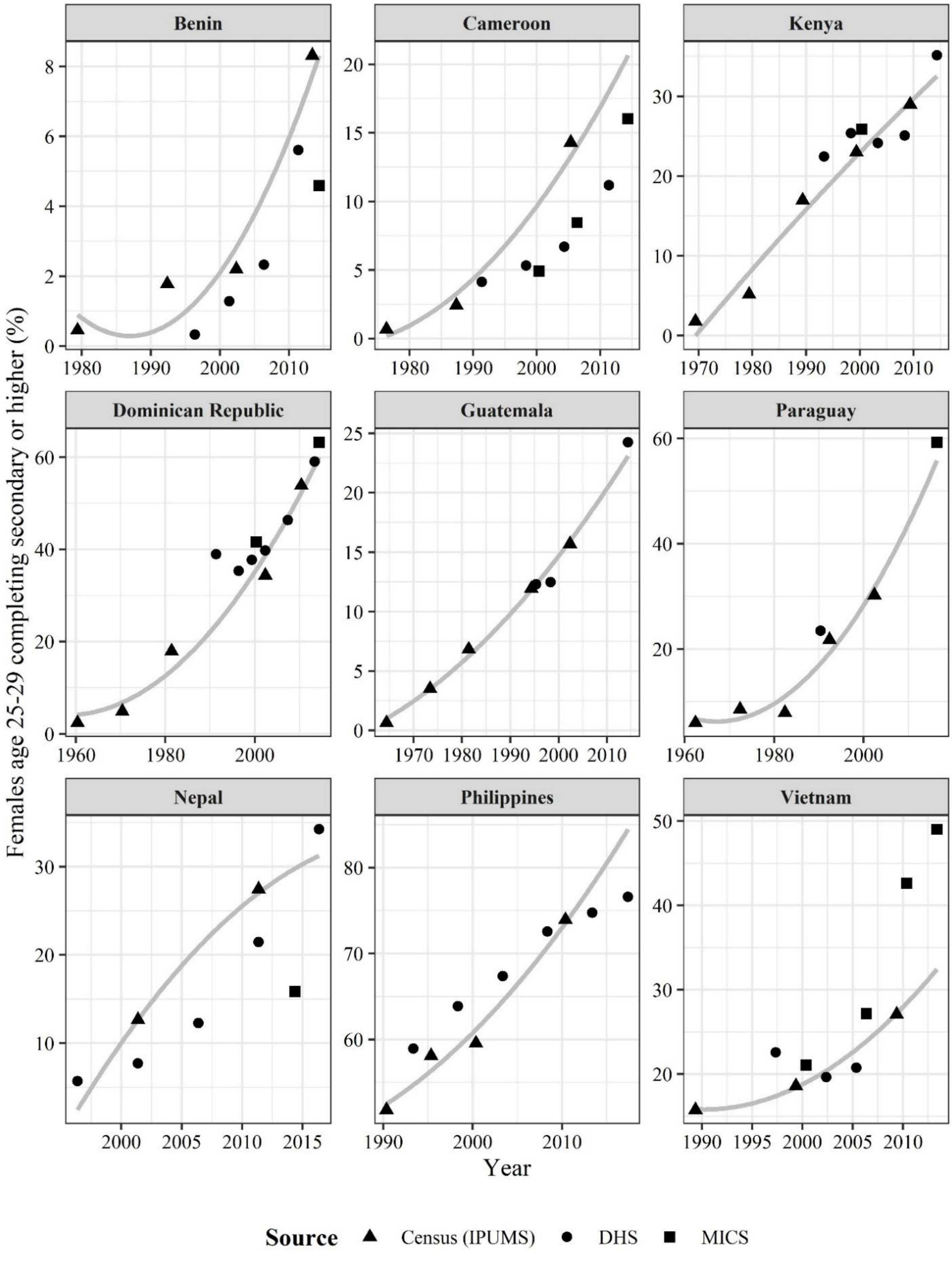
GENERATING NEW TRENDS: COUNTRY EXAMPLES

To address incoherent trends over time, we calculate adjusted estimates of female educational attainment. Providing adjusted estimates poses some challenges. The application of inferential models to correct biases due to poor data quality and to fill in missing values are common place in international research databases (Bauer 2012 et al; Barro and Lee 2013; Cohen and Soto 2007; Kc, Barakat, Goujon, Skirbekk, Sanderson, and Lutz 2010; De la Fuente and Doménech 2002; Jordá and Alonso 2017), and not only for data on education but also for databases measuring mortality, fertility and various economic indicators. The variety of methods used to carry out these estimations makes it difficult to pool or compare data across sources. A parsimonious yet efficient approach is to infer trends from as many observations as possible within a country. Thanks to the availability of international census and survey microdata, we can now perform such an analysis. In our analysis, we use weighted regression to infer trends in educational expansion based on observations from the same country.

Figure 1.2 presents observed point estimates of educational attainment among young females for nine illustrative countries. For each of these countries, at least 8 observations from censuses, DHS surveys, and/or MICS surveys are available. All nine countries experienced an increase in the proportion of females age 25 to 29 completing secondary education or higher between the earliest and most recent census or survey observations, but no country experienced a monotonic increase during the observed period. In Benin, for example, 1.8 percent of females age 25 to 29 had completed secondary or higher according to the 1992 census but only 0.3 percent of females age 25 to 29 had completed secondary or higher according to the 1996 DHS survey. In Nepal, 27.5 percent of females age 25 to 29 had completed secondary or higher according to the 2011 census, but only 15.9 percent in 2014 according to that year's MICS survey.

To identify which observations diverge from overall country-level trends and to what extent, we specify polynomial regression models for each country that estimate the best-fitting trend line according to the survey and census observations, weighted by sample size. As the dependent variable, we use the logit of the proportion of women age 25 to 29 with secondary or more education (ln(P/1-P)) to account for non-linearity in educational expansion. The grey regression line in each country graph depicts the fitted relationship between time and the percent of females age 25 to 29 completing secondary or higher. By design, regression lines favor census observations, which are based on sample sizes much larger than those provided by DHS and MICS surveys. We use regression parameters to produce predicted values of the percentage of females age 25 to 29 completing secondary education or higher that align with expected trends (results in Appendix table A1.1).

Figure 1.2 Percent of women age 25-29 completing secondary education or higher in nine countries

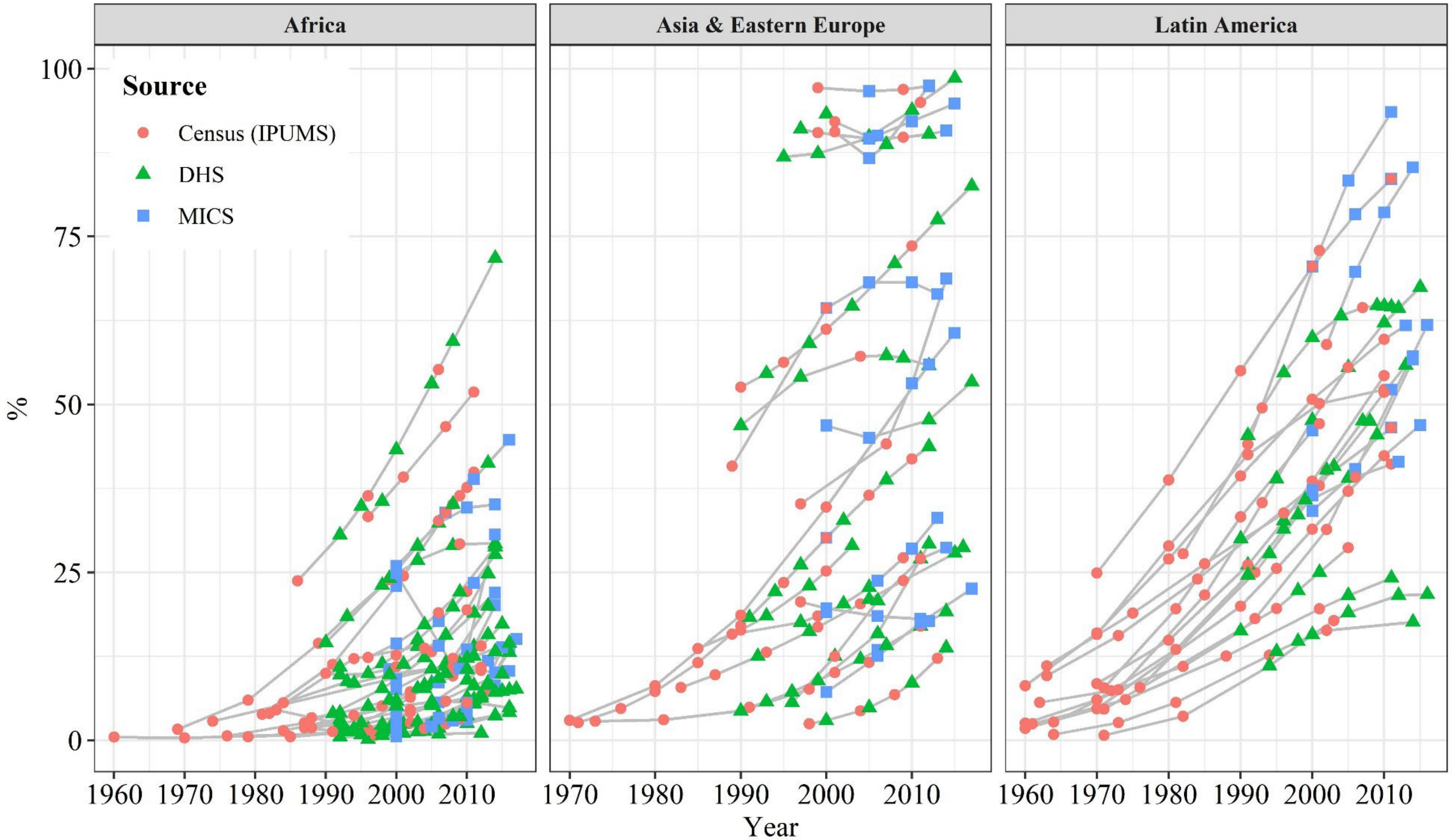


Source Authors' elaboration based on data from IPUMS International, IPUMS DHS, DHS Program, and MICS

ADJUSTED TRENDS

Figure 1.3 plots predicted values of educational attainment among young females for all countries studied. Adjusted data show a monotonically increasing pattern of educational expansion in nearly every country. The few exceptions might reflect actual population trends or might lack of observations, in particular from censuses. To assess coherence in the measurement of female educational attainment across censuses, DHS surveys, and MICS surveys, we produced Figure 1.4 and Table 1.2. Figure 1.4 plots observed and predicted values of the percentage of women age 25 to 29 completing secondary education or higher for all countries studied. Table 1.2 summarizes the absolute difference between observed and predicted values by data source, region, and time period. Overall, coherence across data sources is strong: the mean absolute difference between observed and predicted values is 2.2 percentage points. Of 535 census and survey observations, two-thirds deviate from expected trends by less than two percentage points. Only 11 percent of observations deviate from expected trends by five or more percentage points. The difference between observed and predicted values is 10 or more in 19 of 535 observations (labeled in figure 1.4). As expected, census-based estimates deviate least from expected trends. At the global level, predicted values for DHS-based estimates deviate from observed values by 2.7 percentage points on average compared to 3.2 percentage points for MICS-based estimates. MICS and DHS are similar in terms of the number of observations diverging from predicted estimates by 0 to 2 percentage points, but the plurality of observations with large deviations—10 percentage points or more—are derived from MICS data. In relative terms, the percent of observations deviating more than 10 percentage points from predicted values for MICS is more than twice (8.5 percent) the number than for DHS (3.2 percent).

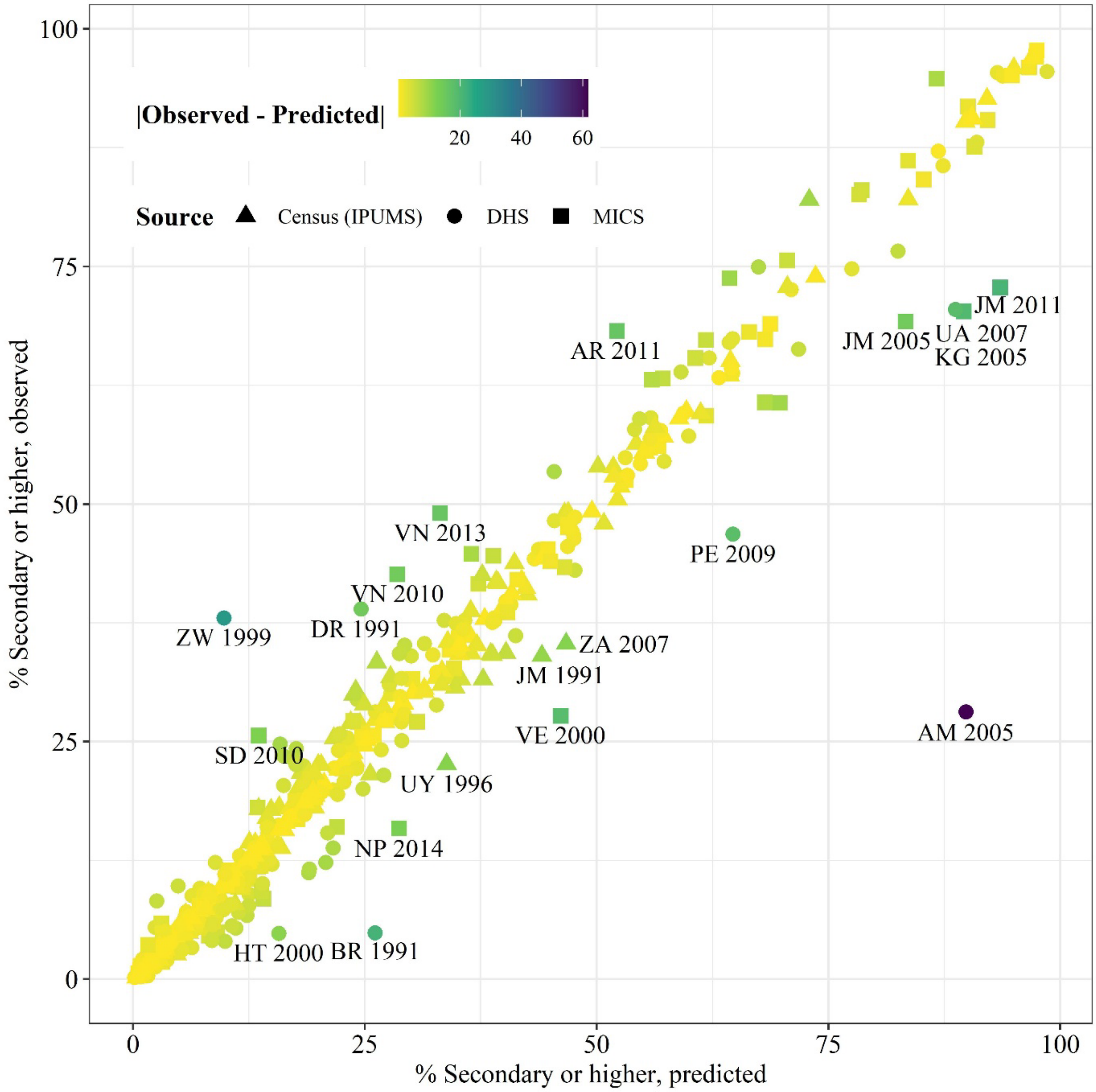
Figure 1.3 Predicted percent of women age 25-29 completing secondary education or higher by year, country, data source and world region



Source Authors' elaboration based on data from IPUMS International, IPUMS DHS, DHS Program, and MICS

In terms of regional differences, Africa and Asia (which includes the few Eastern Europe countries examined in this study) show the highest levels of conformity between observed and expected patterns and across data sources. In Africa, nearly 80 percent of observed point estimates deviate from predicted values by less than 2 percentage points. The share of African observations that diverge 10 or more percentage points is 2 percent or less for all data sources. Compared to MICS-based observations in other regions, MICS surveys in Africa are more in line with expected country trends. In Asia, 70 percent of observations fall within the minimum deviation category (0 to 2 percentage points). MICS surveys in Asia show higher levels of deviation than DHS surveys in the region. Finally, coherence across data sources in Latin America are lowest than in other continent. More than half of all survey and census observations from the region deviate from predicted values by 2 or more percentage points. Results for Latin America are influenced by the predominance of census samples for the region compared to surveys. There are three times as many census samples as DHS surveys and four times as many census samples than MICS surveys available for Latin America. The relative quantity of large census samples means the model demands higher precision from both census and survey observations for most Latin American countries. Accordingly, survey observations perform particularly poorly in the region. Thirty percent of DHS observations and more than 40 percent of MICS observations deviate from predicted values by 5 percentage points or more.

Figure 1.4 Percent of women age 25 to 29 completing secondary or higher, observed versus predicted values



Source Authors' elaboration based on data from IPUMS International, IPUMS DHS, DHS Program, and MICS

Our results do not reveal significant differences in the coherence of observations from the year 2000 or earlier compared to observations after 2000. We also considered the temporal proximity of surveys to censuses. There is no discernible relationship between the number of years elapsed since the previous census and the magnitude of the difference between observed and predicted values for survey observations (Pearson correlation coefficient = -0.06).

Table 1.2 Distribution of samples based on gap between the observed and predicted percentages of women 25-29 with secondary or more

	0-1.9	2-4.9	5-9.9	10+	N
All Countries	67.9	20.7	7.9	3.6	535
Census (IPUMS)	76.2	19.5	2.9	1.4	210
DHS	61.6	24.2	11	3.2	219
MICS	64.2	16	11.3	8.5	106
Africa	79.1	14.8	4.9	1.2	244
Census (IPUMS)	91.7	6.9	0	1.4	72
DHS	71.9	19.8	7.4	1	121
MICS	78.4	13.7	5.9	2	51
Asia & Eastern Europe	69.9	17.8	8.2	4.1	146
Census (IPUMS)	87.3	10.9	1.8	0	55
DHS	56.1	28.1	12.3	3.5	57
MICS	64.7	11.8	11.8	11.8	34
Latin America	46.9	33.8	12.4	6.9	145
Census (IPUMS)	55.4	36.1	6	2.4	83
DHS	39	31.7	19.5	9.8	41
MICS	28.6	28.6	23.8	19	21
2000 or earlier	68.3	23.8	5	2.9	240
Census (IPUMS)	75.4	20.8	2.3	1.5	130
DHS	54.9	32.9	7.3	4.9	82
MICS	75	10.7	10.7	3.6	28
2001 or later	67.5	18.3	10.2	4.1	295
Census (IPUMS)	77.5	17.5	3.8	1.3	80
DHS	65.7	19	13.1	2.2	137
MICS	60.3	17.9	11.5	10.3	78

Note Samples are classified by data source, region, and time period.
Source Authors' elaboration

DISCUSSION

Improved access to census and survey microdata for LMICs has opened new opportunities for large-scale comparative investigations based on multiple sources of data. Used together, these data sources provide more comprehensive temporal and geographic coverage than any single source used in isolation. Yet this type of comparative work is still rare in socio-demographic analysis. Researchers tend to work with a single data source, shying away from the challenges of harmonization and coherence across sources. In this paper, we challenge perceived limitations of cross-national research to examine the measurement of educational expansion across data sources. Literature from a variety of disciplines documents the role of education in personal and social development. Many studies considering this relationship use data from a single country; those studies that do offer a cross-national perspective typically use only one of the three data sources utilized in this paper: censuses, DHS or MICS. As a result, our understanding of the links between education and individual, household, and societal outcomes has been informed by studies using different datasets. Whether these datasets yield similar results was previously unknown. To address this gap, we have carried out a simple coherence exercise to measure the consistency of trends in female educational expansion across data sources. We focus on women age 25 to 29 completing secondary education because gender gaps persist in post-primary education in many LMICs. Analyses of men and other categories of educational attainment produced similar results.

We pooled nearly 20 million individual-level observations from 535 censuses and surveys and 75 countries from IPUMS, DHS and MICS. To maximize comparability across data sources, we use a dichotomous measure of education to identify females completing upper secondary education or higher. These data confirm previously observed patterns of educational expansion: access to education is increasing in all regions, but levels of educational attainment among young women remain low in the majority of LMICs. Less than half of women age 25 to 29 had completed secondary education or higher in the most recent survey or census available in 51 of 75 countries studied. Our coherence analysis, however, shows that the general upward trend in educational attainment among young females conceals erratic trajectories in the majority of countries studied. Our model smooths these trajectories. To assess the accuracy of each point estimate, we compared observed estimates to predicted values. Overall, coherence across data sources is high. Two-thirds of observed point estimates differ from predicted point estimates by two or fewer percentage points. Still, thirty percent of observed point estimates diverge from predicted values by two or more percentage points. Censuses perform better than DHS and MICS. Coherence across data sources is strongest in Africa. Coherence across data sources is lowest in Latin America, where nearly 20 percent of observed point estimates deviate from predicted values by 5 or more percentage points.

Discrepancies across data sources may originate during survey design data collection and/or data prcessing. While DHS and MICS use similar sampling strategies and typically rely on recent censuses as sampling frames, practices vary across countries to adapt to national circumstances. In cases where surveys are fielded soon after the most recent national census, an updated census-based sampling frame may not yet be available, producing discrepancies between data sources even when collected during the same year. For example, in Nepal, the

2001 DHS survey uses the 1991 census as its sampling frame, and 2011 DHS survey uses the 2001 census (though there was a census fielded in 2011). This may explain gaps between DHS- and census-based estimates for these years in Nepal (see figure 1.2). Non-response and other non-sampling errors occur in both censuses and surveys, but there may be systematic differences by data source. For example, refusal to participate may contribute more to non-response for surveys than for censuses which are obligatory or well-promoted in many countries. Likewise, depending on the cultural context, the reporting of educational attainment among female household members might vary depending on the characteristics of interviewers and perceived use of collected information which is likely to differ by data source. As indicated in previous studies (Bauer et al. 2012, Speringer et al. 2019), discrepancies related to data processing are particularly likely with education data due to varied treatment of complete versus incomplete levels of education. In countries where secondary completion rates are still low, the distinction between secondary attendance and completion may not be reported or recorded consistently across sources. Our results identify the countries for which further investigation of these sources of discrepancies may be required for analyses that combine census and survey data.

Results presented here have other practical applications. Adjusted estimates of the percent of women age 25 to 29 completing at least secondary education can be used in a variety of demographic and economic investigations. The data also serve as benchmarks for assessing the accuracy of estimates reported in international databases, the trend alignment of future DHS, MICS, and census-based estimates, and the quality of education information obtained in other sources not included in our analysis such as household surveys with small sample sizes. In future research, we will evaluate consistency across data sources based on other dimensions captured in censuses and household surveys such as demographic composition of the population, family structure, and living arrangements. When data are scarce, as is the case for many LMICs, pooling data resources can be a useful strategy. Still, in the interest of reliable empirical evidence for policymaking, we should not take for granted the coherence of data within and across sources. Our analysis provides empirical validity for most analyses that combine census, DHS and MICS and identifies survey and census samples that should be used with caution.

APPENDIX 1

Table A1.1 Percent of women age 25-29 completing secondary education or higher (P) by country, year, and data source: observed and predicted values

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
Argentina	1970	Census[3]	18.10	15.76	2.34	17162
	1980	Census	28.58	28.94	0.36	100336
	1991	Census	40.47	42.52	2.05	153473
	2001	Census	53.92	50.14	3.78	136778
	2010	Census	50.47	52.24	1.77	156901
	2011	MICS	68.20	52.20	16.00	3317
Armenia	2000	DHS	95.39	93.24	2.15	849
	2001	Census	92.63	92.10	0.53	11471
	2005	DHS	28.13	89.83	61.70	953
	2010	DHS	95.02	93.79	1.23	943
	2011	Census	95.87	94.98	0.89	13888
	2015	DHS	95.51	98.57	3.06	1111
Bangladesh	1991	Census	4.89	4.89	0	471616
	1993	DHS (IPUMS)	7.08	5.75	1.33	2208
	1996	DHS (IPUMS)	9.56	7.22	2.34	2166
	1999	DHS (IPUMS)	12.28	8.89	3.39	2371
	2001	Census	10.03	10.10	0.07	602539
	2004	DHS (IPUMS)	12.55	12.06	0.49	2303
	2006	MICS	18.06	13.44	4.62	12378
	2007	DHS (IPUMS)	14.49	14.14	0.35	2252
	2011	DHS (IPUMS)	17.67	17.01	0.66	3851
	2011	Census	16.96	17.01	0.05	362601
	2012	MICS	17.27	17.74	0.47	10765
	2014	DHS (IPUMS)	19.57	19.16	0.41	3771
Belarus	1999	Census	97.16	97.16	0	34310
	2005	MICS	95.93	96.64	0.71	1252
	2009	Census	96.94	96.91	0.03	36923
	2012	MICS	97.08	97.38	0.30	1358
Benin	1979	Census	0.45	0.52	0.07	15000
	1992	Census	1.78	1.18	0.60	21115
	1996	DHS (IPUMS)	0.33	1.59	1.26	1092
	2001	DHS (IPUMS)	1.28	2.41	1.13	1314
	2002	Census	2.20	2.63	0.43	29520
	2006	DHS (IPUMS)	2.33	3.77	1.44	4079
	2011	DHS (IPUMS)	5.61	6.08	0.47	3417
	2013	Census	8.32	7.42	0.90	42883
	2014	MICS	4.59	8.20	3.61	2859
Bolivia	1976	Census	7.91	7.88	0.03	17702
	1992	Census	24.56	24.98	0.42	24420
	1994	DHS	29.98	27.79	2.19	1496
	1998	DHS	37.75	33.57	4.18	1834

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
	2000	MICS	44.74	36.48	8.26	771
	2001	Census	37.91	37.92	0.01	31187
	2003	DHS	39.42	40.77	1.35	2877
	2008	DHS	46.69	47.52	0.83	2928
Botswana	1981	Census	4.21	3.88	0.33	3696
	1991	Census	9.67	11.30	1.63	5470
	2000	MICS	23.22	22.96	0.26	1188
	2001	Census	26.86	24.47	2.39	7424
	2011	Census	39.18	39.90	0.72	10408
Brazil	1970	Census	7.25	8.43	1.18	175630
	1980	Census	17.83	14.89	2.94	242723
	1991	DHS	4.87	26.12	21.25	1108
	1991	Census	27.25	26.12	1.13	362413
	1996	DHS	28.84	32.73	3.89	2291
	2000	Census	34.21	38.57	4.36	407987
	2010	Census	56.35	54.27	2.08	416354
Burkina Faso	1985	Census	0.54	0.53	0.01	33016
	1993	DHS (IPUMS)	1.25	1.28	0.03	1296
	1996	Census	1.62	1.67	0.05	38956
	1998	DHS (IPUMS)	1.41	1.95	0.54	1171
	2003	DHS (IPUMS)	2.23	2.71	0.48	2145
	2006	MICS	1.81	3.16	1.35	1451
	2006	Census	3.40	3.16	0.24	56225
	2010	DHS (IPUMS)	2.01	3.68	1.67	3064
Burundi	2000	MICS	3.63	3.16	0.47	721
	2005	MICS	1.96	2.38	0.42	1501
	2010	DHS (IPUMS)	2.94	2.51	0.43	1689
	2016	DHS (IPUMS)	4.02	4.12	0.10	3023
Cambodia	1998	Census	2.48	2.46	0.02	45814
	2000	DHS	2.98	2.96	0.02	2193
	2004	Survey[4] (IPUMS)	3.37	4.40	1.03	2990
	2005	DHS	3.58	4.89	1.31	2320
	2008	Survey[4] (IPUMS)	6.96	6.80	0.16	62643
	2010	DHS	8.60	8.56	0.04	3396
	2013	Census	12.29	12.21	0.08	6015
	2014	DHS	12.28	13.77	1.49	3111
Cameroon	1976	Census	0.69	0.66	0.03	27308
	1987	Census	2.43	2.60	0.17	34447
	1991	DHS (IPUMS)	4.14	3.99	0.15	754
	1998	DHS (IPUMS)	5.33	7.71	2.38	1033
	2000	MICS	4.92	9.10	4.18	975
	2004	DHS (IPUMS)	6.70	12.29	5.59	1933
	2005	Census	14.31	13.17	1.14	72900
	2006	MICS	8.45	14.08	5.63	1595
	2011	DHS (IPUMS)	11.18	18.94	7.76	2859
	2014	MICS	16.02	22.00	5.98	1854

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
Central African Rep.	1994	DHS (IPUMS)	1.72	1.77	0.05	1098
	2000	MICS	2.39	2.32	0.07	3255
	2006	MICS	3.16	3.34	0.18	2257
	2010	MICS	4.62	4.51	0.11	2290
Chad	1996	DHS (IPUMS)	0.17	0.18	0.01	1480
	2000	MICS	0.62	0.56	0.06	1130
	2004	DHS (IPUMS)	1.34	1.33	0.01	1179
	2010	MICS	2.88	2.99	0.11	3305
	2014	DHS (IPUMS)	3.76	3.69	0.07	3579
Colombia	1964	Census	2.05	2.72	0.67	12319
	1973	Census	7.48	7.55	0.07	70229
	1985	Census	25.36	21.66	3.70	122574
	1990	DHS	33.99	30.00	3.99	1786
	1993	Census	31.57	35.35	3.78	152653
	1995	DHS	37.99	38.94	0.95	2045
	2000	DHS	43.02	47.67	4.65	1919
	2005	DHS	56.03	55.52	0.51	6363
	2005	Census	55.75	55.52	0.23	153211
	2010	DHS	65.37	62.15	3.22	8165
	2015	DHS	74.93	67.46	7.47	6322
Congo	2005	DHS (IPUMS)	8.47	8.49	0.02	1399
	2009	DHS (IPUMS)	11.55	11.36	0.19	1237
	2011	DHS (IPUMS)	13.80	13.94	0.14	2101
	2014	MICS	20.20	20.14	0.06	1926
Congo DR	2000	MICS	7.53	7.64	0.11	2188
	2007	DHS (IPUMS)	12.96	11.50	1.46	1782
	2010	MICS	11.86	13.50	1.64	2299
	2013	DHS (IPUMS)	16.20	15.71	0.49	3590
Costa Rica	1963	Census	8.34	9.65	1.31	2641
	1973	Census	14.16	15.60	1.44	6271
	1984	Census	30.36	24.00	6.36	10629
	2000	Census	31.57	37.75	6.18	15029
	2011	MICS	43.35	46.54	3.19	985
	2011	Census	49.01	46.54	2.47	19652
Cote d'Ivoire	1994	DHS (IPUMS)	0.86	1.37	0.51	1534
	1998	DHS (IPUMS)	5.46	2.40	3.06	569
	2000	MICS	5.87	3.07	2.80	1926
	2005	DHS (IPUMS)	3.63	5.18	1.55	1067
	2006	MICS	3.83	5.66	1.83	2268
	2011	DHS (IPUMS)	8.12	8.15	0.03	2123
	2016	MICS	11.46	10.34	1.12	2325
Cuba	2002	Census	59.00	58.91	0.09	43868
	2006	MICS	60.62	69.74	9.12	1274
	2010	MICS	83.03	78.59	4.44	1791
	2014	MICS	84.15	85.28	1.13	1961
Dominican Republic	1960	Census	2.37	2.55	0.18	7222

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
	1970	Census	4.88	6.01	1.13	9222
	1981	Census	17.92	13.53	4.39	17021
	1991	DHS	38.93	24.62	14.31	1424
	1996	DHS	35.32	31.44	3.88	1564
	1999	DHS	37.75	35.78	1.97	224
	2000	MICS	41.60	37.25	4.35	653
	2002	Census	34.35	40.21	5.86	35614
	2002	DHS	39.78	40.21	0.43	3894
	2007	DHS	46.35	47.54	1.19	4434
	2010	Census	53.86	51.79	2.07	38209
	2013	DHS	59.06	55.85	3.21	1474
	2014	MICS	63.22	57.15	6.07	5545
Egypt	1986	Census	23.83	23.76	0.07	261427
	1992	DHS (IPUMS)	30.48	30.58	0.10	2419
	1995	DHS (IPUMS)	37.42	34.85	2.57	3293
	1996	Census	36.03	36.41	0.38	226212
	2000	DHS (IPUMS)	44.21	43.29	0.92	3424
	2005	DHS (IPUMS)	54.86	53.11	1.75	4658
	2006	Census	55.44	55.19	0.25	323772
	2008	DHS (IPUMS)	59.54	59.39	0.15	4029
	2014	DHS (IPUMS)	66.28	71.79	5.51	5593
Eswatini	2000	MICS	25.62	25.94	0.32	882
	2006	DHS (IPUMS)	34.12	32.33	1.79	809
	2010	MICS	32.68	34.63	1.95	897
	2014	MICS	35.91	35.11	0.80	803
Ethiopia	1984	Census	1.41	1.44	0.03	117671
	1994	Census	3.81	3.70	0.11	192189
	2000	DHS (IPUMS)	5.21	5.08	0.13	2907
	2005	DHS (IPUMS)	5.94	5.74	0.20	2791
	2007	Census	5.20	5.82	0.62	55486
	2011	DHS (IPUMS)	6.59	5.65	0.94	3490
	2016	DHS (IPUMS)	9.79	4.86	4.93	3185
Ghana	1984	Census	5.60	5.58	0.02	54221
	1993	DHS (IPUMS)	4.43	8.72	4.29	903
	1998	DHS (IPUMS)	7.03	11.38	4.35	940
	2000	Census	12.92	12.69	0.23	78720
	2003	DHS (IPUMS)	12.09	14.99	2.90	1058
	2006	MICS	16.81	17.73	0.92	1019
	2008	DHS (IPUMS)	20.18	19.83	0.35	1968
	2010	Census	22.06	22.17	0.11	110632
	2011	MICS	24.68	23.44	1.24	1701
	2014	DHS (IPUMS)	30.98	27.64	3.34	1678
Guatemala	1964	Census	0.63	0.88	0.25	7460
	1973	Census	3.50	2.64	0.86	10147
	1981	Census	6.83	5.65	1.18	11083
	1994	Census	11.94	12.69	0.75	29376

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
	1995	DHS	12.30	13.22	0.92	1947
	1998	DHS	12.49	14.72	2.23	1026
	2002	Census	15.70	16.35	0.65	41847
	2014	DHS	24.25	17.61	6.64	4206
Guinea	1983	Census	4.53	4.52	0.01	20630
	1996	Census	1.52	1.54	0.02	31379
	1999	DHS (IPUMS)	2.44	1.60	0.84	1383
	2005	DHS (IPUMS)	1.75	2.38	0.63	1332
	2012	DHS (IPUMS)	8.77	6.37	2.40	1677
	2016	MICS	11.94	13.79	1.85	1964
Guyana	2000	MICS	34.65	34.16	0.49	814
	2005	DHS	37.73	39.02	1.29	418
	2006	MICS	38.59	40.41	1.82	769
	2009	DHS	48.24	45.46	2.78	763
	2014	MICS	56.12	56.66	0.54	976
Haiti	1971	Census	0.76	0.74	0.02	16771
	1982	Census	3.17	3.56	0.39	6648
	1994	DHS	5.37	11.10	5.73	902
	2000	DHS	4.80	15.73	10.93	1589
	2003	Census	20.08	17.81	2.27	36205
	2005	DHS	11.59	19.02	7.43	1769
	2012	DHS	13.80	21.60	7.80	2403
	2016	DHS	22.22	21.72	0.50	2256
Honduras	1961	Census	3.21	2.44	0.77	700
	1974	Census	5.40	6.08	0.68	9166
	1988	Census	14.39	12.54	1.85	15661
	2001	Census	18.07	19.58	1.51	22690
	2005	DHS	19.91	21.58	1.67	3566
	2011	DHS	29.42	24.17	5.25	3981
India	1983	Survey[5] (IPUMS)	8.72	7.86	0.86	24520
	1987	Survey[5] (IPUMS)	9.64	9.77	0.13	27059
	1992	DHS (IPUMS)	8.92	12.51	3.59	22344
	1993	Survey[5] (IPUMS)	13.01	13.10	0.09	23612
	1998	DHS (IPUMS)	20.38	16.24	4.14	22949
	1999	Survey[5] (IPUMS)	17.98	16.90	1.08	24926
	2004	Survey[5] (IPUMS)	22.54	20.30	2.24	24527
	2005	DHS (IPUMS)	15.39	20.99	5.60	23759
	2009	Survey[5] (IPUMS)	29.96	23.78	6.18	19333
	2015	DHS (IPUMS)	27.50	27.88	0.38	124239
Indonesia	1971	Census	2.65	2.62	0.03	26274
	1976	Survey[6] (IPUMS)	4.86	4.74	0.12	9953
	1980	Census	7.28	7.25	0.03	283374
	1985	Survey[6] (IPUMS)	9.51	11.54	2.03	26639
	1990	Census	17.10	17.07	0.03	41494
	1991	DHS	18.78	18.29	0.49	5832
	1994	DHS	24.07	22.14	1.93	6739

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
	1995	Survey[6] (IPUMS)	27.03	23.47	3.56	31556
	1997	DHS	28.12	26.15	1.97	6860
	2000	MICS	31.59	30.15	1.44	1834
	2000	Census	30.13	30.15	0.02	951041
	2002	DHS	32.27	32.76	0.49	6744
	2005	Survey[6] (IPUMS)	34.27	36.47	2.20	48676
	2007	DHS	37.52	38.77	1.25	7956
	2010	Census	41.94	41.90	0.04	1059239
	2012	DHS	45.21	43.74	1.47	7682
Iraq	1997	Census	20.61	20.60	0.01	79448
	2000	MICS	19.12	19.68	0.56	3759
	2006	MICS	18.82	18.49	0.33	4452
	2011	MICS	18.04	18.10	0.06	9025
Jamaica	1982	Census	31.80	27.76	4.04	8181
	1991	Census	34.04	44.08	10.04	10401
	2001	Census	82.02	72.92	9.10	8461
	2005	MICS	69.20	83.32	14.12	492
	2011	MICS	72.77	93.53	20.76	736
Jordan	1990	DHS	45.51	46.88	1.37	2089
	1997	DHS	57.83	54.09	3.74	1971
	2004	Census	57.12	57.18	0.06	21753
	2007	DHS	54.49	57.27	2.78	3167
	2009	DHS	57.72	56.92	0.80	2874
	2012	DHS	56.84	55.77	1.07	3145
Kazakhstan	1995	DHS	87.13	86.84	0.29	592
	1999	DHS	85.58	87.36	1.78	734
	2006	MICS	91.83	90.05	1.78	1943
	2010	MICS	90.39	92.17	1.78	2059
	2015	MICS	95.10	94.81	0.29	2215
Kenya	1969	Census	1.79	1.66	0.13	26558
	1979	Census	5.18	5.98	0.80	45921
	1989	Census	16.97	14.43	2.54	42454
	1993	DHS (IPUMS)	22.44	18.41	4.03	1338
	1998	DHS (IPUMS)	25.40	23.09	2.31	1412
	1999	Census	22.99	23.92	0.93	60253
	2000	MICS	25.87	24.72	1.15	1853
	2003	DHS (IPUMS)	24.13	26.78	2.65	1539
	2008	DHS (IPUMS)	25.09	28.98	3.89	1497
	2009	Census	28.99	29.22	0.23	167262
	2014	DHS (IPUMS)	35.13	29.30	5.83	6334
Kyrgyz Republic	1997	DHS	88.07	91.01	2.94	598
	1999	Census	90.72	90.46	0.26	18992
	2005	MICS	70.27	89.60	19.33	1019
	2009	Census	90.24	89.76	0.48	23637
	2012	DHS	90.65	90.26	0.39	1331
	2014	MICS	87.59	90.75	3.16	1191

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
Laos	2000	MICS	7.70	7.20	0.50	1488
	2005	Census	11.54	11.59	0.05	21626
	2006	MICS	10.57	12.56	1.99	1157
	2011	MICS	18.75	17.50	1.25	3884
	2017	MICS	22.18	22.58	0.40	4260
Lesotho	1996	Census	12.34	12.34	0	6976
	2000	MICS	13.82	14.42	0.60	1102
	2004	DHS (IPUMS)	17.21	17.22	0.01	1456
	2006	Census	19.49	18.96	0.53	8026
	2009	DHS (IPUMS)	19.44	22.07	2.63	1789
	2014	DHS (IPUMS)	29.72	28.82	0.90	1572
Liberia	1974	Census	2.83	2.83	0	6725
	2007	DHS (IPUMS)	11.24	11.10	0.14	1296
	2008	Census	12.18	12.20	0.02	15151
	2013	DHS (IPUMS)	20.04	20.01	0.03	1718
Malawi	1987	Census	1.85	1.85	0	30849
	1992	DHS (IPUMS)	2.76	3.06	0.30	887
	1998	Census	5.09	5.10	0.01	39520
	2000	DHS (IPUMS)	6.52	5.92	0.60	2479
	2004	DHS (IPUMS)	8.62	7.69	0.93	2301
	2006	MICS	6.07	8.63	2.56	5007
	2008	Census	9.87	9.58	0.29	57404
	2010	DHS (IPUMS)	10.94	10.52	0.42	4538
	2013	MICS	9.97	11.87	1.90	4502
	2016	DHS (IPUMS)	13.84	13.11	0.73	4225
Mali	1987	Census	1.98	1.89	0.09	30491
	1995	DHS (IPUMS)	0.27	0.90	0.63	1903
	1998	Census	0.84	0.89	0.05	35398
	2001	DHS (IPUMS)	2.10	1.04	1.06	2492
	2006	DHS (IPUMS)	1.26	1.86	0.60	2774
	2009	MICS	4.83	3.21	1.62	4935
	2009	Census	3.42	3.21	0.21	54723
	2012	DHS (IPUMS)	3.32	6.34	3.02	2335
	2015	MICS	9.24	13.88	4.64	3523
Mexico	1960	Census	1.70	1.77	0.07	18860
	1970	Census	2.61	4.71	2.10	17035
	1990	Census	22.55	19.96	2.59	337284
	1995	Census	21.58	25.58	4.00	13721
	2000	Census	30.57	31.41	0.84	430071
	2005	Census	35.15	37.10	1.95	440673
	2010	Census	41.19	42.35	1.16	458103
	2015	Survey[7] (IPUMS)	49.08	46.94	2.14	440085
	2015	MICS	47.80	46.94	0.86	2290
Mongolia	1989	Census	40.60	40.81	0.21	8623
	2000	MICS	73.75	64.35	9.40	1863
	2000	Census	63.52	64.35	0.83	10967

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
	2005	MICS	60.68	68.14	7.46	1387
	2010	MICS	67.31	68.19	0.88	1489
	2013	MICS	68.10	66.46	1.64	2103
Morocco	1982	Census	4.01	4.01	0	38493
	1992	DHS (IPUMS)	10.10	10.92	0.82	1562
	1994	Census	12.22	12.16	0.06	53354
	2003	DHS (IPUMS)	10.06	13.97	3.91	2852
	2004	Census	13.88	13.71	0.17	64611
Mozambique	1997	DHS (IPUMS)	0.25	0.76	0.51	1887
	1997	Census	0.78	0.76	0.02	65059
	2003	DHS (IPUMS)	1.47	1.32	0.15	2542
	2007	Census	2.42	2.44	0.02	83801
	2008	MICS	3.50	2.93	0.57	2882
	2009	DHS (IPUMS)	3.30	3.56	0.26	1076
	2011	DHS (IPUMS)	5.52	5.41	0.11	2424
Nepal	1996	DHS	5.69	5.64	0.05	1838
	2001	DHS	7.72	12.49	4.77	1863
	2001	Census	12.66	12.49	0.17	82787
	2006	DHS	12.27	20.78	8.51	1841
	2011	DHS	21.47	27.05	5.58	2219
	2011	Census	27.45	27.05	0.40	142388
	2014	MICS	15.86	28.69	12.83	2461
	2016	DHS	34.24	28.73	5.51	2210
Nicaragua	1971	Census	4.68	4.67	0.01	6485
	1995	Census	19.29	19.64	0.35	17001
	1998	DHS	25.82	22.29	3.53	2367
	2001	DHS	25.41	25.01	0.40	2245
	2005	Census	28.55	28.68	0.13	21272
Niger	1992	DHS (IPUMS)	0.45	0.49	0.04	1443
	1998	DHS (IPUMS)	0.79	0.69	0.10	1388
	2000	MICS	0.97	0.76	0.21	996
	2006	DHS (IPUMS)	0.74	0.95	0.21	1847
	2012	DHS (IPUMS)	1.17	1.09	0.08	2432
Nigeria	1990	DHS (IPUMS)	15.29	14.55	0.74	1836
	1999	DHS (IPUMS)	22.16	24.11	1.95	1637
	2003	DHS (IPUMS)	27.06	28.95	1.89	1471
	2006	Survey[8] (IPUMS)	31.73	32.68	0.95	3674
	2007	MICS	31.87	33.93	2.06	5394
	2007	Survey[8] (IPUMS)	35.49	33.93	1.56	3679
	2008	Survey[8] (IPUMS)	34.19	35.17	0.98	4748
	2008	DHS (IPUMS)	34.67	35.17	0.50	6815
	2009	Survey[8] (IPUMS)	38.74	36.41	2.33	3385
	2010	Survey[8] (IPUMS)	42.51	37.64	4.87	3121
	2011	MICS	44.54	38.87	5.67	6342
	2013	DHS (IPUMS)	36.15	41.27	5.12	7381
	2016	MICS	45.24	44.75	0.49	6201

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
Pakistan	1973	Census	2.10	2.87	0.77	49036
	1981	Census	3.34	3.06	0.28	340212
	1990	DHS (IPUMS)	4.92	4.37	0.55	1974
	1998	Census	7.15	7.60	0.45	469545
	2006	DHS (IPUMS)	24.72	15.88	8.84	29428
	2012	DHS (IPUMS)	27.87	29.24	1.37	3982
Palestine	1997	Census	35.19	35.20	0.01	9101
	2007	Census	44.28	44.12	0.16	8145
	2010	MICS	52.50	53.14	0.64	2743
	2014	MICS	68.96	68.72	0.24	2065
Panama	1960	Census	9.12	8.11	1.01	1926
	1970	Census	13.81	15.97	2.16	5335
	1980	Census	28.66	26.98	1.68	7740
	1990	Census	41.70	39.37	2.33	10055
	2000	Census	47.93	50.77	2.84	12097
	2010	Census	59.78	59.67	0.11	13446
	2013	MICS	67.27	61.78	5.49	1497
Paraguay	1962	Census	5.95	5.65	0.30	2896
	1972	Census	8.53	7.40	1.13	7511
	1982	Census	7.89	11.00	3.11	11577
	1990	DHS	23.46	16.28	7.18	1054
	1992	Census	21.74	18.11	3.63	15747
	2002	Census	30.22	31.40	1.18	18117
	2016	MICS	59.26	61.82	2.56	1372
Peru	1991	DHS	53.41	45.41	8.00	2848
	1993	Census	49.22	49.49	0.27	92731
	1996	DHS	54.24	54.72	0.48	5309
	2000	DHS	57.16	59.94	2.78	4644
	2004	DHS	63.28	63.19	0.09	6546
	2007	Census	65.06	64.43	0.63	116384
	2009	DHS	46.83	64.70	17.87	3792
	2010	DHS	63.75	64.68	0.93	3576
	2011	DHS	64.27	64.55	0.28	3502
	2012	DHS	67.00	64.30	2.70	3650
Philippines	1990	Census	51.82	52.56	0.74	249175
	1993	DHS	58.95	54.64	4.31	2614
	1995	Census	58.12	56.28	1.84	286916
	1998	DHS	63.88	59.11	4.77	2420
	2000	Census	59.56	61.22	1.66	292832
	2003	DHS	67.38	64.67	2.71	2221
	2008	DHS	72.56	70.98	1.58	2279
	2010	Census	73.94	73.60	0.34	375739
	2013	DHS	74.74	77.52	2.78	2461
	2017	DHS	76.59	82.51	5.92	4133
Rwanda	1992	DHS (IPUMS)	1.56	1.71	0.15	1131
	2000	DHS (IPUMS)	4.14	3.47	0.67	1647

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
	2000	MICS	2.88	3.47	0.59	739
	2002	Census	4.21	4.17	0.04	30290
	2005	DHS (IPUMS)	5.28	5.54	0.26	1820
	2010	DHS (IPUMS)	6.48	8.96	2.48	2567
	2012	Census	11.01	10.88	0.13	46952
	2014	DHS (IPUMS)	13.47	13.21	0.26	2352
Senegal	1988	Census	1.85	1.87	0.02	28965
	1992	DHS (IPUMS)	3.09	2.52	0.57	1181
	2000	MICS	3.31	4.13	0.82	2434
	2002	Census	4.87	4.58	0.29	38821
	2005	DHS (IPUMS)	3.06	5.26	2.20	2646
	2010	DHS (IPUMS)	5.59	6.37	0.78	2974
	2012	DHS (IPUMS)	5.30	6.77	1.47	1648
	2014	DHS (IPUMS)	5.99	7.15	1.16	1624
	2015	DHS (IPUMS)	7.98	7.33	0.65	1713
	2016	DHS (IPUMS)	8.00	7.49	0.51	1657
	2017	DHS (IPUMS)	9.06	7.65	1.41	3016
Sierra Leone	2000	MICS	1.41	0.82	0.59	980
	2004	Census	1.63	1.73	0.10	22401
	2005	MICS	2.22	2.08	0.14	2070
	2008	DHS (IPUMS)	5.04	3.55	1.49	1811
	2010	MICS	5.01	5.01	0	2679
	2013	DHS (IPUMS)	9.04	8.21	0.83	2891
	2017	MICS	14.05	15.09	1.04	3103
South Africa	1996	Census	32.41	33.31	0.90	158301
	1998	DHS (IPUMS)	36.84	35.61	1.23	1907
	2001	Census	41.74	39.19	2.55	166246
	2007	Survey[9] (IPUMS)	35.31	46.71	11.40	42443
	2011	Census	52.91	51.86	1.05	209390
Sudan	2000	MICS	25.35	24.83	0.52	6951
	2008	Census	10.95	11.05	0.10	212414
	2010	MICS	25.62	13.55	12.07	3444
	2014	MICS	27.07	30.63	3.56	3718
Tajikistan	2000	MICS	47.57	46.88	0.69	936
	2005	MICS	43.98	45.02	1.04	1503
	2012	DHS	48.61	47.67	0.94	1676
	2017	DHS	53.00	53.33	0.33	1938
Tanzania	1988	Census	3.36	3.36	0	98157
	1992	DHS (IPUMS)	3.85	4.09	0.24	1737
	1996	DHS (IPUMS)	5.65	5.07	0.58	1554
	1999	DHS (IPUMS)	6.59	6.03	0.56	818
	2002	Census	7.24	7.23	0.01	165233
	2003	DHS (IPUMS)	7.4	7.7	0.30	1365
	2004	DHS (IPUMS)	9.3	8.21	1.09	2009
	2007	DHS (IPUMS)	10.02	9.99	0.03	1697
	2010	DHS (IPUMS)	11.18	12.23	1.05	1774

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
	2012	Census	13.97	114.04	0.07	183634
	2015	DHS (IPUMS)	23.1	17.32	5.78	2307
Thailand	1970	Census	2.83	2.97	0.14	27739
	1980	Census	8.92	8.12	0.80	15679
	1990	Census	20.67	18.62	2.05	24278
	2000	Census	30.70	34.73	4.03	27316
	2012	MICS	63.07	55.96	7.11	3072
	2015	MICS	65.36	60.65	4.71	3625
Togo	1960	Census	0.99	0.49	0.50	805
	1970	Census	0.17	0.37	0.20	1181
	1998	DHS (IPUMS)	1.19	1.31	0.12	1751
	2000	MICS	3.60	1.61	1.99	960
	2006	MICS	3.36	3.27	0.09	1275
	2010	MICS	4.85	5.60	0.75	1242
	2010	Census	5.73	5.60	0.13	26741
	2013	DHS (IPUMS)	5.71	8.60	2.89	1768
Trinidad and Tobago	1970	Census	28.90	24.92	3.98	2115
	1980	Census	34.12	38.76	4.64	4336
	1990	Census	55.40	55.02	0.38	5262
	2000	MICS	75.63	70.56	5.07	527
	2000	Census	72.84	70.56	2.28	4263
	2006	MICS	82.56	78.31	4.25	666
	2011	MICS	86.14	83.60	2.54	756
	2011	Census	82.07	83.60	1.53	5449
Turkey	1985	Census	13.64	13.66	0.02	98502
	1990	Census	16.46	16.40	0.06	118632
	1993	DHS	17.32	18.52	1.20	1577
	1998	DHS	21.70	23.01	1.31	1587
	2000	Census	25.16	25.20	0.04	146356
	2003	DHS	31.63	28.96	2.67	2043
Uganda	1991	Census	1.28	1.30	0.02	63199
	1995	DHS (IPUMS)	8.21	2.58	5.63	1462
	2000	DHS (IPUMS)	5.69	5.14	0.55	1494
	2002	Census	6.40	6.43	0.03	94313
	2006	DHS (IPUMS)	7.20	9.25	2.05	1649
	2011	DHS (IPUMS)	10.63	12.52	1.89	1792
	2016	DHS (IPUMS)	15.98	14.47	1.51	3392
Ukraine	2001	Census	90.55	90.56	0.01	170044
	2005	MICS	94.75	86.67	8.08	1358
	2007	DHS	70.48	88.71	18.23	1087
	2012	MICS	97.72	97.48	0.24	1985
Uruguay	1963	Census	9.68	11.07	1.39	9449
	1975	Census	21.63	18.93	2.70	9735
	1985	Census	33.35	26.29	7.06	10917
	1996	Census	22.62	33.83	11.21	10898
	2006	Survey[10] (IPUMS)	41.85	39.18	2.67	8300

Country	Year	Data Source	Observed P	Predicted P	Difference[1]	Sample size[2]
	2011	Census	43.76	41.13	2.63	11590
	2012	MICS	42.03	41.45	0.58	478
Venezuela	1971	Census	7.38	7.82	0.44	39993
	1981	Census	21.66	19.57	2.09	62069
	1990	Census	30.99	33.28	2.29	71554
	2000	MICS	27.70	46.14	18.44	745
	2001	Census	47.81	47.11	0.70	96038
Vietnam	1989	Census	15.75	15.79	0.04	125526
	1997	DHS	22.58	17.58	5.00	1336
	1999	Census	18.59	18.50	0.09	102748
	2000	MICS	21.05	19.05	2.00	1336
	2002	DHS	19.65	20.33	0.68	1256
	2005	DHS	20.75	22.77	2.02	992
	2006	MICS	27.15	23.74	3.41	1245
	2009	Census	27.07	27.16	0.09	621601
	2010	MICS	42.60	28.50	14.10	1853
	2013	MICS	49.02	33.13	15.89	1423
Zambia	1990	Census	10.24	9.99	0.25	29844
	1992	DHS (IPUMS)	7.31	9.77	2.46	1285
	1996	DHS (IPUMS)	3.98	9.92	5.94	1427
	1999	MICS	10.51	10.60	0.09	1780
	2000	Census	11.22	10.95	0.27	37791
	2001	DHS (IPUMS)	10.08	11.36	1.28	1465
	2007	DHS (IPUMS)	13.96	15.65	1.69	1479
	2010	Census	19.77	19.42	0.35	54052
	2013	DHS (IPUMS)	20.01	24.80	4.79	3049
Zimbabwe	1994	DHS (IPUMS)	4.07	8.50	4.43	1026
	1999	DHS (IPUMS)	38.01	9.81	28.20	1088
	2005	DHS (IPUMS)	5.61	10.66	5.05	1664
	2009	MICS	10.57	10.68	0.11	2238
	2010	DHS (IPUMS)	7.85	10.62	2.77	1858
	2012	Census	10.58	10.41	0.17	30674
	2014	MICS	9.96	10.10	0.14	2519
	2015	DHS (IPUMS)	11.04	9.91	1.13	1800

[1] Absolute value of observed value less predicted value
[2] Number of females age 25 to 29 (unweighted)
[3] All census samples accessed from IPUMS International
[4] Cambodia Intercensal Population Survey
[5] National Sample Survey Organization Socio-Economic Survey of India
[6] Indonesia Intercensal Population Survey
[7] Intercensal Survey
[8] Nigeria: National Bureau of Statistics General Household Survey
[9] South Africa Community Survey
[10] Uruguay Extended National Survey of Homes 2006
Source Authors' elaboration

CHAPTER 2

DISCREPANCIES IN HOUSEHOLD SIZE AND COMPOSITION BETWEEN CENSUS AND DHS DATA ACROSS 22 LOW- AND MIDDLE-INCOME COUNTRIES

INTRODUCTION

Population and housing censuses and Demographic and Health Surveys (DHS) represent two of the most widely used data sources for demographic research. Thanks to the availability of harmonized microdata from the DHS Program (ICF 2025; Boyle et al. 2024) and IPUMS International (Ruggles et al. 2025), researchers are increasingly designing comparative studies to explore trends over time and across countries and regions. To maximize geographic and temporal scope, comparative research on this scale often relies on evidence from both these data infrastructures, using available samples to paint the most comprehensive picture possible of trends across countries and over time. Large scale cross-national and cross-temporal studies provide important insights into, for example, the differential impacts of national economic, social, political, and environmental conditions on demographic outcomes, because they allow researchers to study variation in these processes across a variety of contexts.

Despite the importance of and increase in these types of large-scale comparative demographic studies (see for example Ackah et al. 2025; Batyra and Pesando 2021; Esteve et al. 2017; Esteve et al. 2020; Esteve and Reher 2021; Esteve and Reher 2024), research assessing the validity and reliability of results based on data pooled across sources is limited. Population censuses and DHS surveys are both household-based enumerations, making the selection and coverage of households a fundamental element of resulting data quality. Discrepancies in estimates of household size and composition may point to systematic differences in the populations being captured across the two data sources with implications for demographic studies on nearly every other topic. Nonetheless, existing studies investigating consistency and comparability of DHS surveys and other data sources have largely overlooked this topic, focusing instead on fertility and mortality estimates. To fill this gap, this study undertakes an exploratory analysis of coherence across DHS survey data and census samples available IPUMS International. We use aggregate and individual-level data from 78 census samples and 123 DHS surveys collected during the period 1969 to 2019 to assess coherence in the measurement of household size and composition across data sources in 22 low-and middle-income countries (LMICs).

This paper documents trends in average household size over time across the two data sources and identifies the sources of discrepancies in household size. Differences in average household size suggest that census samples and DHS surveys may be systematically capturing households with distinct structures. We find that average household size is often larger in DHS samples than in census samples, driven by differences across the two data sources in the average number of children per household. Women with children are more likely to be captured in DHS samples than in census samples, and in select countries this is particularly true for higher parity women. Our study provides practical evidence for the research community about the validity of comparative investigations using two important data sources for demographic studies.

BACKGROUND

Population and housing censuses are generally regarded as the most comprehensive source of population data for LMICs. In the absence of population registers, which are still rare in parts of the world, censuses represent the only source of information on the entire population in many countries. For this reason, censuses are a foundational source of data for policymaking and for demographic investigations, especially those concerned with the size, distribution, and characteristics of a country's population. Many high-income countries have introduced register-based and combined methods that make partial or full use of administrative data to collect census information. In most low- and middle-income countries, however, censuses still involve a full-field enumeration that aims to collect information from each household in a country on a range of topics at a specified time (United Nations 2025). While questionnaire type (single, short, long) and data collection approach (face-to-face or telephone interviews, self-enumeration online or using mail-in questionnaires) may vary across contexts, traditional censuses typically rely on a common methodology that locates each dwelling in a country and counts all present individuals and, in some cases, absent individuals who are usual residents of the household.

The main advantage of a traditional census is its comprehensive coverage of the population. On the other hand, censuses are complex, costly operations and, as a result, are usually conducted only once every 10 years. Smaller in scale, sample surveys require fewer resources, are conducted more frequently, and cover a broader range of topics in greater depth. The Demographic and Health Surveys (DHS) are nationally representative household surveys that collect data on a number of topics related to population, health and nutrition (ICF 2019). DHS surveys are conducted about every 5 years. Since 1984, more than 400 DHS surveys have been completed in over 90 countries in Africa, Latin America, Asia, Oceania and Eastern Europe.

Generally representative at the national level, DHS surveys are usually based on a stratified two-stage cluster design. In the first stage, enumeration areas are drawn from census files or an alternative sampling frame, and in the second stage a sample of households is drawn in each selected enumeration area (ICF International 2012). This type of probability sampling relies on updated and reliable sampling frames. Censuses are generally considered the most suitable sampling frames for DHS surveys. For this reason, we expect strong coherence between census-based estimates and survey-based estimates. When a census has not been conducted recently, alternative sampling frames such as electoral rosters are used. While DHS surveys target women of child-bearing age and their young children, and sufficient precision of indicators on these populations is used to determine sample sizes, all households selected in each enumeration area are surveyed regardless of the presence of these target groups.

Both censuses and DHS surveys allow for the study of household and family characteristics, including size and structure. Traditional censuses assign members to household based on usual place of residence and record the relationship between each household member and the household reference person. DHS surveys collect basic demographic information on all usual residents and visitors in all surveyed households, making it a suitable source for nationally representative information on household size and structure.

POTENTIAL SOURCES OF ERROR IN DHS SURVEYS AND CENSUSES

While both data sources are designed to be nationally representative, differences in their design and implementation have the potential to produce biased statistics and estimates. Inaccuracies in statistical data can generally be separated into two categories: **sampling errors** and **non-sampling errors**.

Sampling errors occur due to discrepancies between the sample population and the true population. Sampling error occurs with all sample surveys and is an important source of variation between survey-based estimates and other data sources (Pullum, Assaf, and Staveteig 2017). Surveys reflect only one of many different samples that could have been selected, each yielding slightly different results. The extent to which sample-based estimates diverge with those for the whole population depends on the sample design and size. While DHS surveys generally follow a standard model, sample design can vary significantly across countries (Verma and Lê 1996). DHS calculates and provides sampling weights that aim to minimize the impacts of sampling error.

Conversely, censuses aim to enumerate the entire population of a country, and estimates based on complete-count census data are unlikely to suffer from significant sampling error. Oftentimes, demographic research relies on public- or scientific-use samples of census microdata like those disseminated by IPUMS. While these samples are systematically drawn from complete-count census data to provide representative data on the entire population, *post facto* sampling error can nonetheless occur. Most IPUMS samples are simple random samples that do not require differential weights. For samples with documented coverage or non-response issues, differential weights are provided.

Both censuses and surveys are subject to **non-sampling errors** which are inaccuracies that occur during data collection or processing that are not related to sample design. There are various types of non-sampling errors, including those related to specification, coverage, non-response, measurement, and processing (Brieumer and Lyberg 2003).

When non-sampling errors occur randomly across large data collection exercises, their impacts on resulting national-level estimates are limited. Errors that occur systematically, however, lead to biased estimates. Non-sampling error is very difficult to measure and resulting biases may be hard to identify. While methods exist to produce estimates of non-response and certain measurement errors (age-heaping, for example), the required metadata and reference datasets from post-enumeration surveys or population registers are not always available. Furthermore, given the wide variety of estimates and analyses produced, it is almost impossible to determine the impacts of these errors across all potential results, especially those at the sub-national level. Analysis of coherence across data sources, such as that carried out in this paper, can also provide evidence around potential non-sampling errors.

QUALITY OF DHS SURVEY AND CENSUS DATA

The quality of DHS survey data has been assessed along several different dimensions including accuracy of birth estimates (Pullum and Becker 2014; Schoumaker 2014) and fertility and maternal mortality data (Ahmed et al. 2014; Schoumaker 2014; Pullum, Assaf, and Staveteig 2017), quality and consistency of age and date reporting (Lyons-Amos and Stones 2017; Pullum and Staveteig 2017; Singh, Kashyap and Bango 2022), and potential interviewer (Pullum et al. 2018; Elkasabi and Khan 2022) and questionnaire design effects (Courtney, Fleuret, and Ahmed 2020). Results from these methodological studies indicate acceptable data quality in most surveys but also identify several potential sources of non-sampling error in survey fieldwork.

Across DHS surveys, birth histories provided by female respondents are subject to the omission of births and deaths and the misreporting of date of birth and death (Schoumaker 2014; Pullum and Becker 2014). Respondents and interviewers may omit recent births or misreport the date of birth of a child (birth displacement) to avoid completing lengthy survey sections on child health for children under age five (Schoumaker 2011; Schoumaker 2014). The omission of distant births, especially for children who have left home or died, is also common in birth histories (Schoumaker 2014). Likewise, some respondents may fail to include children who have died in birth histories or accurately report child deaths (Pullum and Becker 2014), with direct impacts on estimates of household size. Based on an analysis of 192 DHS surveys, Pullum and Becker (2014) estimate that the mean probability of omitting a birth is 1.8 percent and that the mean probability of displacing a birth is 1.5 percent. The estimated mean probability of omitting a birth that resulted in death or an under-five death is higher—5.3 percent. The omission of recent births and the misreporting of dates of births impacts the quality of fertility estimates in DHS surveys (Schoumaker 2014). Age misreporting has been documented for adult respondents as well, in the form of incomplete information on date of birth, age heaping, and age displacement to avoid eligibility for further interviewing (Lyons-Amos and Stones 2017; Pullum and Staveteig 2017; Singh, Kashyap and Bango 2022).

Interviewer characteristics also impact data quality, with older and more highly educated interviewers as well as those with previous experience with DHS or survey fieldwork collecting higher quality data (Pullum et al. 2018). These interviewer effects may be exacerbated by long and non-factual questions and questions on complex or difficult topics (Elkasabi and Azam 2022). The relationship between questionnaire design and data quality is somewhat unclear. In an analysis of the determinants of birth history omissions in 52 DHS surveys, Schoumaker (2011) documents correlations between omissions and the length of the survey and the length of reference period of the child health module. On the other hand, in an examination of questionnaire length and data quality in three DHS surveys, Allen, Fleuret and Ahmed (2020) find no evidence to support the idea that longer questionnaires result in data quality issues or that interviewers may intentionally reduce their workload.

The quality of census data is generally evaluated according to its coverage. In many countries, estimates of coverage or undercounts are generated based on post-enumeration surveys or demographic analysis of vital records or population registers. In many LMICs, however, this type of systematic analysis of census coverage is not carried out. While some of the non-sampling errors and data quality issues documented in DHS surveys like age heaping also apply to census fieldwork, evaluations of census data quality along these lines are comparatively limited. Some country- and census-specific analyses of age-data quality exist for LMICs (Jowett and Li 1992; Mukherjee and Mukhopadhyay 1988; Nagi, Stockwell and Snavley 1973; Shipanga and Shinyemba 2023; Stockwell and Dixon 1966, Yadav, Vishwakarma and Chauhan 2020), though many are from several decades ago or focus on historical census data. The United Nations (see United Nations 2026) and IPUMS International (see IPUMS International 2026a) have produced cross-national assessments of the quality of age and sex data collected in censuses which indicate significant variation across countries and censuses. At the same time, census questionnaires tend to be shorter with simpler questions, which may contribute to better data quality overall. As censuses generally do not have age-based modules and questionnaires, enumerators and respondents may have less reason to misreport age or other information.

The enumeration approach of the census may also impact the results of some analyses. In many countries, the census captures both the *de facto* (present in household at the time of the census) and the *de jure* (usual resident of the household at the time of the census) population, distinguishing between usual residents who are present, usual residents who are absent, and visitors. In some countries, however, only a *de facto* or a *de jure* approach is used. In countries employing the *de facto* approach, usual household residents who are away from home at the time of the census are not captured, with implications for the measurement of household size and structure (Ackah et al. 2025). Likewise, visitors may be included in measures of household size and structure. DHS surveys generally take a *de facto* approach, administering questionnaires to all eligible persons in the household at the time of the survey regardless of whether he or she is a household resident or visitor.

COHERENCE ACROSS DATA SOURCES

Existing studies that investigate data quality of DHS surveys and census tend to focus on statistical accuracy and coherence over time within each data source. Studies that examine coherence of estimates across data sources focus on a specific topic or subset of indicators including age at first union (Hertrich, Lardoux and George 2014), educational attainment (Jeffers and Esteve 2021), fertility and mortality (Pullum, Assaf, Staveteig 2017; Singh et al 2022) and seasonal bias in data collection (Wright, Yang and Walker 2012), on a subpopulation such as older Africans (Randall and Coast 2016), or on a single country— Gambia (Reremoi et al 2019), Ghana (Ackah et al 2025), and Malawi (Tasciotti and Wagner 2018). Findings from these studies highlight the importance of case-by-case analysis of the coherence and comparability of microdata across sources (Tasciotti and Wagner 2018). The assessment of comparability across data sources is necessary as researchers are increasingly combining data from censuses and various surveys in cross-national studies across demographic topics including demographic and family change (Furstenberg 2019; Tabutin et al 2020); union- and family-formation (Batyra and Pesando 2021; Esteve et al. 2017), and household structure and living arrangements (Esteve et al. 2020; Esteve and Reher 2021; Esteve and Reher 2024; Pohl, Esteve, Galeano 2025).

The present analysis examines for the first time the coherence of DHS survey data and census data disseminated by IPUMS International to measure household size and composition. Discrepancies in estimates of household size and composition may point to systematic differences in the populations being captured across the two data sources with implications for demographic studies on nearly every other topic. Information on coherence or divergence across censuses and DHS surveys in measures of household size and structure can shed light on the general suitability of pooling these sources for cross-national demographic research.

In the context of the unique purposes and methodologies of censuses and DHS surveys described above, we address several original research questions: Are estimates of average household size for a given country consistent across census and DHS survey data? What are the potential sources of discrepancies in household size across data sources? The answers to these questions provide valuable information to the research community on the validity of the growing body of demographic research pooling data from censuses and various international survey programs.

DATA AND METHODS

We use aggregate and individual-level data from 78 census samples available from IPUMS International (Ruggles et al. 2025) and 123 DHS surveys (ICF 2004-2017) collected during the period 1969 to 2019 to assess coherence in the measurement of household size and composition across these two data sources in 22 LMICs.

IPUMS census samples vary in size but typically include 10 percent of the total population. While some IPUMS census samples include group quarters and institutional households, only private households are considered in all phases of this analysis. In DHS surveys, only private households are included. Sample sizes for DHS surveys typically cover 20,000 to 50,000 households. Sample size information is available on the IPUMS International website[4] and in the DHS final survey reports[5]. In both IPUMS and DHS microdata files, individuals are organized into households, allowing for the analysis of household structure.

The first stages of the analysis rely on aggregated census and DHS survey data available in the CORESIDENCE Database (CoDB), which provides access to 146 harmonized indicators on household size and composition for 156 countries, 3950 subnational areas, and 58 data points in time (Galeano et al. 2024). Indicators in CoDB have been consistently calculated from individual microdata samples available from IPUMS and the DHS Program. We use the following CoDB variables: average household size, average number of adults, average number of children under age five, average number of children under age 18, average number of male adults, and average number of female adults, and proportion of households by size of household (proportion of one-person households, two-persons households, etc.). Average number of female adults and average number of male adults are calculated using the CoDB variables on average number of male and female adults in male- and female-headed households and the proportion of male- and female-headed households.

The CoDB household size and structure variables are derived based on all individual records included in the sample regardless of household residence status. Some census samples available

[4] See https://international.ipums.org/international-action/sample_details
[5] See https://dhsprogram.com/publications/index.cfm

from IPUMS include a variable which distinguishes between usual residents and visitors, but this variable is not always available. Likewise, DHS surveys generally do not cover usual residents who were absent at the time of the survey. Therefore, it is not possible to consistently count only usual household residents (the *de jure* population) for household size and structure variables across samples. In aggregate, this has little impact on the national level measures presented in CoDB. In *de facto* census and DHS samples, a visitor will be counted in the household he or she is visiting but will not be counted in his or her household of residence, balancing national-level averages. In censuses that employ both a *de facto* and *de jure* approach, a small number of individuals may be counted in both the household they are visiting and in their household of residence.

We begin the analysis by documenting trends in average household size over time across the two data sources. We include in the analysis all countries with three or more annual observations in each data source. Censuses and DHS surveys are rarely conducted during the same year in a given country. To quantify the size of discrepancies in average household size and other household characteristics between data sources, it is necessary to produce estimates of household size that refer to the same year. To achieve this, we perform linear interpolation to estimate annual values for the variables listed in the previous paragraph by country by data source. We interpolate values between empirical observations using the approx() function in R resulting in an annual time series per country per source spanning the earliest empirical observation to the latest empirical observation. We reduce the sample of country-level observations to years for which there is at least one empirical observation across the two data sources, restricting the time span to the period of overlap between the two data sources. For example, for Bangladesh, DHS data are available for the years 1993, 1996, 1999, 2004, 2007, and 2011; census data are available for the years 1991, 2001, and 2011. Linear interpolation provides annual estimates for the period 1993 to 2011 for DHS and 1991-2011 for IPUMS. The years selected for analysis are 1993, 1996, 1999, 2001, 2004, and 2011.

The analysis consists of two main parts. The first phase documents differences in average household size between IPUMS and DHS using the sample of interpolated, country-level data from CoDB. We also begin to identify the potential sources of discrepancies in household size by using linear regression to estimate the relationship between average household size and data source when controlling for other household characteristics. We run fourteen models with average household size as the dependent variable. Model 1 includes data source as the only independent variable. Model 2 adds time period with year grouped in five-year intervals as an additional independent variable, and model 3 adds country. Inconsistencies in household size may point to systematic differences in the types of individuals and households being captured across data sources. To explore that issue, models 4 to 14 include data source, time period, country, and one of the following CoDB variables: average number of adults, average number of children under age five, average number of children under age 18, average number of male adults, and average number of female adults, and proportion of households by size of household (proportion of one-person households, two-persons households, etc.).

Informed by the results of the first phase of analysis that indicate that differences in household size across data sources are related to the number of children in the household, the second phases use census microdata from IPUMS International and DHS survey data accessed from IPUMS DHS (Boyle et al. 2024) to examine the socio-demographic characteristics of women aged 15 to 49 captured in the two data sources. For this analysis, we focus on four African countries where DHS and census data were collected within two years of one another: Benin (DHS, 2011; Census, 2013), Guinea (DHS, 2012; Census, 2014), Senegal (DHS, 2012; Census, 2013), and Tanzania (DHS, 2010; Census, 2012). For the DHS data, we append household size information from the household member DHS files (HR) to the individual women's file (IR) so that women are the unit of analysis.

We first document the absolute percentage point differences between DHS and census-based calculations of the share of women who are single, childless, and who live in one-person, two-person, three-person, etc. households. Using pooled DHS and census microdata, we then use logistic regression to estimate the odds that an individual observation derives from DHS or census when we control for these characteristics. This type of analysis provides evidence on the extent to which individual-level female characteristics predict the source of data, pointing to potential biases in the data source. We run two logistic regression models for each country. The dependent variable in both models is data source. Model 1 considers parity by age only with age group and number of children ever born as the only independent variables. Age is grouped into five-year intervals. Children ever born represents the total number of children ever born to the respondent. In Model 2, we include other individual characteristics as potential determinants of data source. Marital status includes three categories: (1) never married, (2) married, and (3) separated, divorced, or widowed. Educational attainment categorizes women based on highest level of schooling *completed* at the time of the survey or census: (1) less than primary school, (2) primary, (3) secondary or higher. Urban or rural location of the woman's household reflects the country-specific urban-rural definition applied by the DHS or census. We categorize whether the individual's household was located in an urban or rural area according to the DHS and IPUMS variables.

For consistency with the CoDB data used in the first phase of the analysis, household size corresponds to the total number of person records associated with the household record in the sample. The Tanzania 2012 census and DHS surveys for all four countries employed a *de facto* approach, so visitors are included in measures of household size and usual residents absent at the time of data collection may be omitted. The Benin 2013, Guinea 2014 and Senegal 2013 censuses captured both *de facto* and *de jure* residents, though the microdata sample available from IPUMS for Guinea 2014 only includes usual household residents (absent and present). For the Benin 2013 and Senegal 2013 census samples, usual residents and visitors are included in measures of household size. For these samples, results based on the *de jure* population only do not differ from results based on the entire sample. We therefore use all available cases and do not restrict the sample based on residence status. While certain relevant individual characteristics such as marital status and number of children ever born are not collected for household visitors in the Benin, Guinea, and Senegal census samples, these data are collected for absent usual residents and therefore are reflected in the aggregate and individual-level analyses.

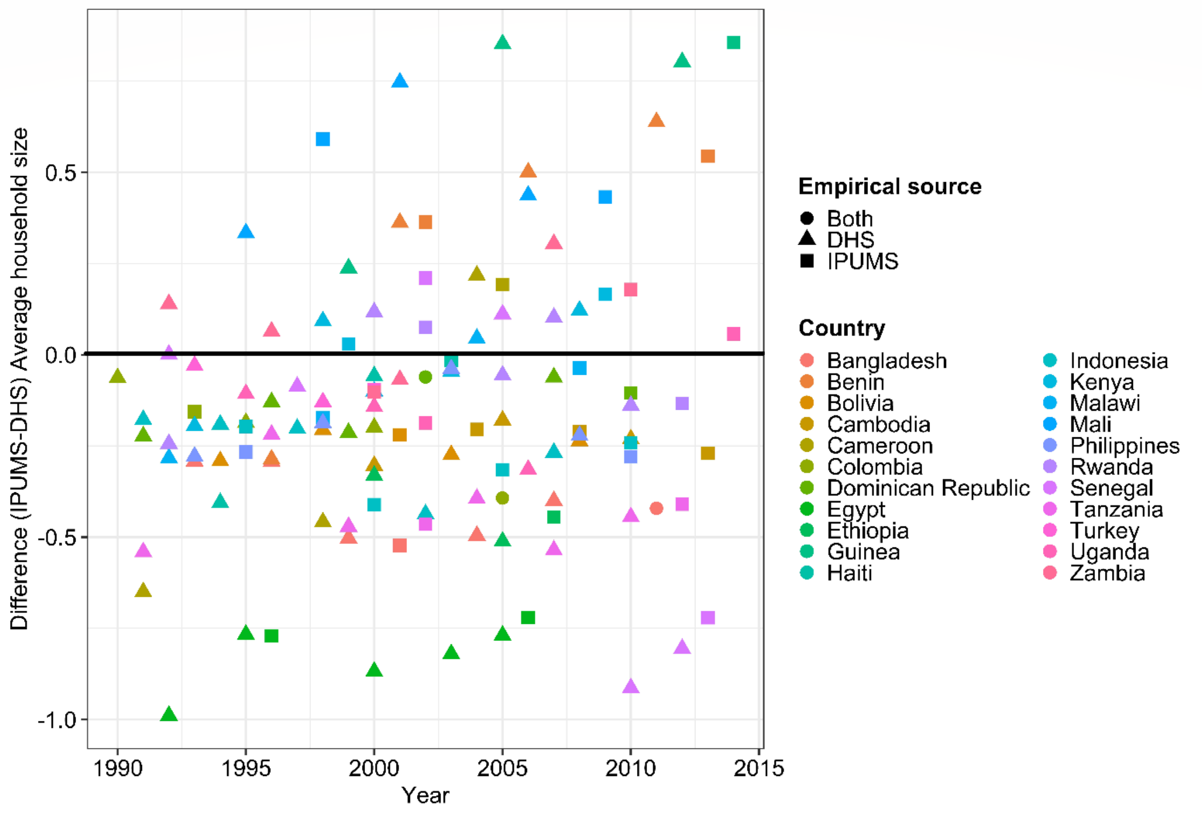
The variables included in the models are highly comparable across the samples pooled for analysis. Basic demographic variables such as age, sex, and marital status are completely comparable across DHS and IPUMS samples for our purposes. While urban-rural definitions vary by country they usually align across data sources within a country. Sampling weights were re-scaled based on sample size in the pooled microdata file.

RESULTS

BIAS AT A GLOBAL SCALE

Figure 2.1 provides a visual representation of the difference in average household size in number of persons between IPUMS census samples and DHS surveys by country and year for 22 countries included in the initial phase of analysis (see also Appendix table A2.1). Results are based on the interpolated data and include years for which there is at least one empirical observation. Each symbol represents a country-year observation. The color of the symbol indicates the country, and the shape of the symbol indicates the data source of the empirical observation: DHS, IPUMS, or both. For country-year observations plotted above the black line at zero, average household size is larger in the IPUMS census sample than in the DHS survey. For observations plotted below the black line at zero, the average household size is smaller in the IPUMS census sample than in the DHS survey. For nearly three-quarters of country-year observations (92 of 124), average household size is smaller in the IPUMS census sample than in DHS. For only 32 country-year observations is average household size larger in the IPUMS census sample than in DHS.

Figure 2.1 Difference in average household size between census samples from IPUMS and DHS surveys in number of persons, by country and year (IPUMS less DHS)



Source Authors' elaboration based on data from CORESIDENCE Database

The mean absolute difference in household size between census and DHS observations is 0.31 persons. The absolute difference is 0.5 persons or higher for 20 percent of observations (25 of 124). There is no clear trend by year of data collection or source of empirical data. Five countries account for the majority of observations with absolute differences of 0.5 persons or more: Benin, Egypt, Guinea, Mali, and Senegal.

Differences in average household size suggest that IPUMS census samples and DHS samples may be systematically capturing households with distinct structures. Smaller average household sizes in IPUMS suggest that smaller households may be underrepresented or larger households may be overrepresented in DHS samples. To begin to explore these discrepancies and identify how structure and composition of households differ between IPUMS and DHS samples, we run a series of linear regression models to estimate the relationship between average household size and data source when controlling for other household characteristics.

ORIGINS OF DISCREPANCIES

The results of the regression models presented in Table 2.1 confirm larger average household size in DHS samples compared to IPUMS census samples and provide evidence around the origins of these discrepancies. Model 1 includes data source as the only independent variable and indicates that average household size is on average 0.15 persons larger in DHS samples compared to IPUMS census samples. The coefficient for data source remains consistent and becomes statistically significant when time period and country are included in the model. Changes to the coefficient for data source when variables measuring household composition are included in the model point to the root of differences in average household size. If the addition of a particular household composition control does not change the size, direction, or significance of the coefficient for data source, it suggests the control does not explain the discrepancy. Alternatively, if the addition of a household composition control does impact the coefficient, it suggests this control may be driving differences in household size.

Table 2.1 Results of linear regression with average household size as dependent variable

Model	Independent variables	**βData source**	Significance
1	Data source (IPUMS as reference)	0.15	
2	Model 1 + Period	0.15	
3	Model 2 + Country	0.15	***
4	Model 3 + Avg. no adults	0.11	***
5	Model 3 + Avg. no. adult males	0.16	***
6	Model 3 + Avg. no. adult females	0.078	***
7	Model 3 + Avg. no. children below age 18	-0.02	
8	Model 3 + Avg. no. children below age 5	0.05	**
9	Model 3 + Share 1 person HHs	0.02	
10	Model 3 + Share 2 person HHs	-0.06	*
11	Model 3 + Share 3 person HHs	-0.03	
12	Model 3 + Share 4 person HHs	0.14	***
13	Model 3 + Share 5 person HHs	0.20	***
14	Model 3 + Share 6+ person HHs	-0.05	**

***p < 0.001 | ** p < 0.01 | * p < 0.05
Source Authors' elaboration based on data from CORESIDENCE Database

When average number of adults in the household is controlled for in model 4, the size of the coefficient for data source decreases only slightly and remains significant, indicating that other differences between data sources explain discrepancies in average household size. This is also the case for model 5 which controls for average number of adults males in the household. The coefficient for data source decreases substantially in magnitude but remains significant when average number of adult females is controlled for in model 6, suggesting this indicator of household composition may explain some of the differences in average household size between data sources. When average number of children under 18 is controlled for in model 7, the magnitude of the coefficient for data source is greatly reduced and becomes insignificant. The direction of the coefficient also changes, suggesting that differences in the average number of children under 18 are driving differences in average household size. Similarly, controlling for the average number of children under age 5 also leads to a decrease in magnitude and significance for the data source coefficient.

When it comes to the distribution of households by size, the results indicate differences in the share of smaller and larger household across data sources. When the share of one-, two-, and three-persons households is controlled for, the magnitude and significance of the data source coefficient is greatly reduced. This effect is also observed when the share of 6-person or larger households is controlled for.

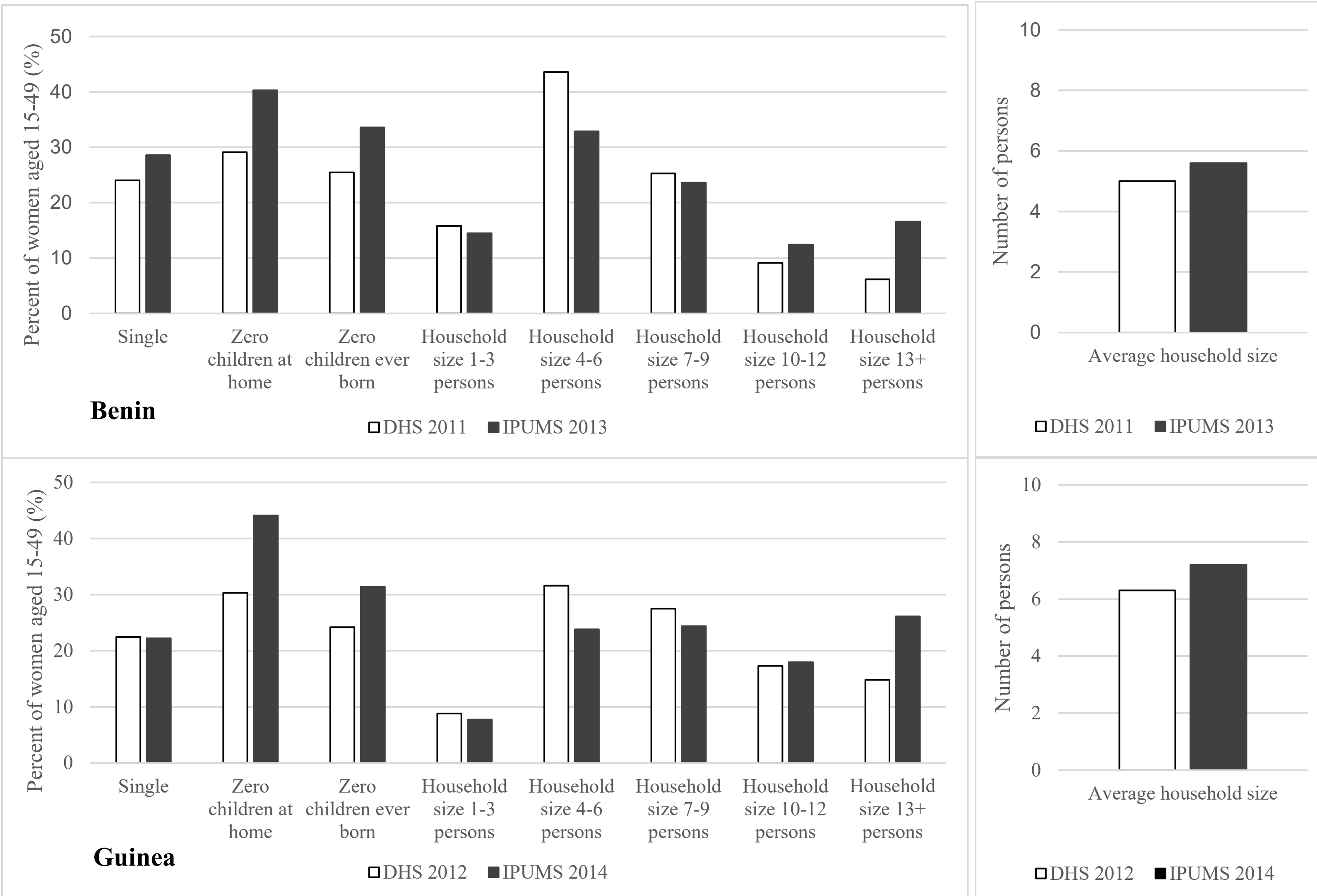
DIFFERENCES AT THE INDIVIDUAL LEVEL

The regression models provide initial evidence around the origins of differences between DHS and IPUMS census samples. The results suggest that DHS surveys often capture households with more children. To validate these findings, we use microdata from IPUMS International and IPUMS DHS for Benin, Guinea, Senegal, and Tanzania to assess more closely the characteristics of women aged 15 to 49 years included in each data source. We select samples for each country collected within two years of one another. In Benin, a DHS survey was fielded in 2011 and the census was conducted in 2013. In Guinea, a DHS was fielded in 2012 and the census was conducted in 2014. In Senegal, a DHS was fielded in 2012 and the census in 2013. In Tanzania, a DHS was fielded in 2010 and the census in 2012. We include both countries that follow the general trend of larger average household size in DHS samples (Senegal and Tanzania) and those where average household size is smaller in DHS census samples (Benin and Guinea) to examine the characteristics of women captured in both situations.

Figure 2.2 presents select indicators describing women aged 15 to 49 and average household size by data source—DHS survey or IPUMS cenus sample. In Benin, Senegal, and Tanzania, the share of single women is higher in the IPUMS sample than in the DHS survey. In Guinea, the share of single women is about the same in both data sources. In all four countries, the share of women with no children living at home and the share of women with zero children ever born is higher in the IPUMS sample than in the DHS survey. Gaps between data sources for no children living at home are significant, ranging from 8.5 percentage points in Senegal to nearly 14 percentage points in Guinea, and are larger than gaps for zero children ever born. Sizeable gaps in the share of women with no children living at home and zero children ever born provide further evidence that DHS surveys tend to capture women with children more often than censuses.

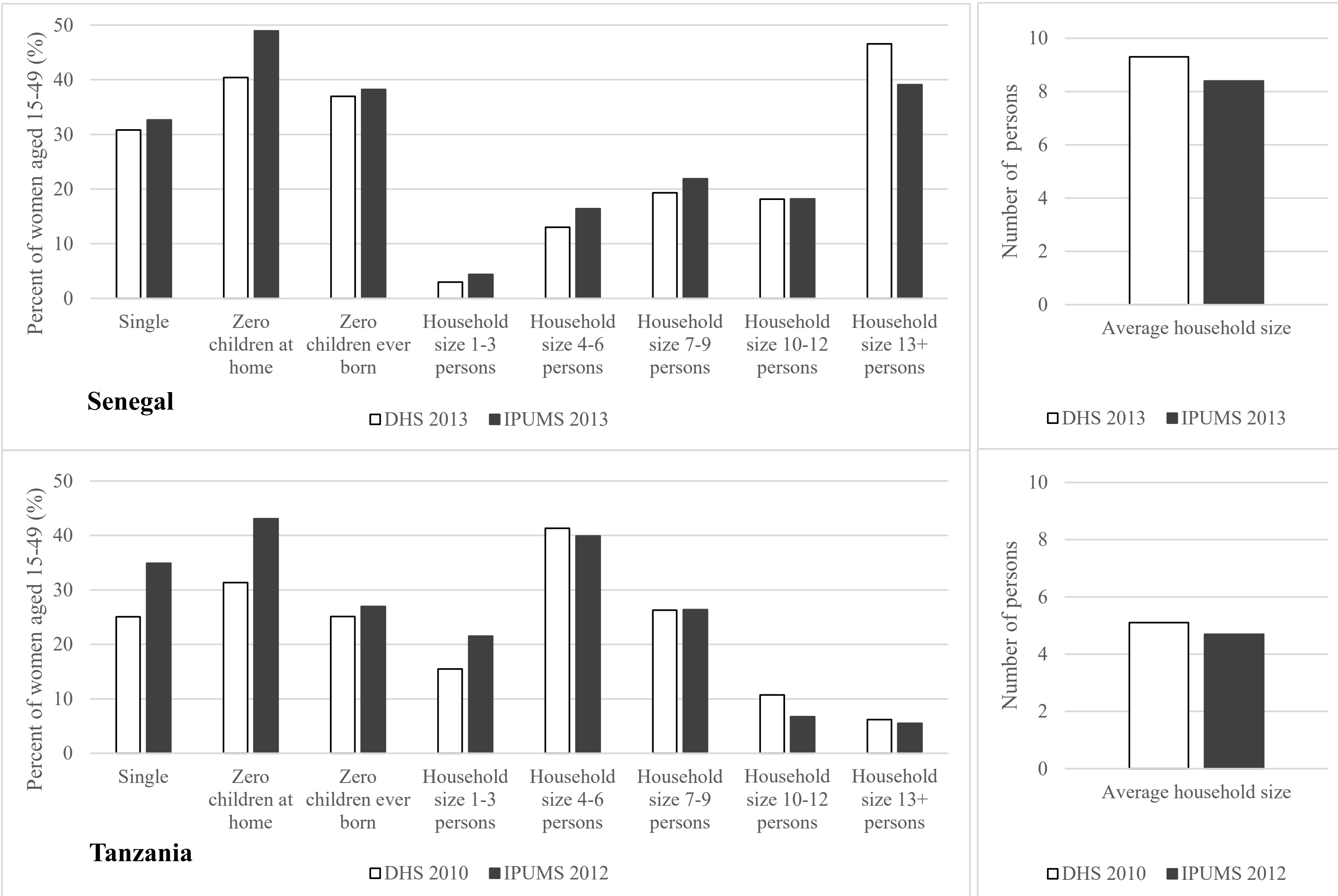
In Benin and Guinea—where average household size is smaller in DHS than in census samples—the share of women living in small households of one to three persons is higher in the DHS survey than in the IPUMS sample, though differences are small. The share of women living in mid-size households (four to nine persons) is also higher in the DHS survey than in the IPUMS census sample, with larger discrepancies between data sources compared to gaps for small households. In these two countries, the share of women in households of 10 to 12 persons is lower in the DHS survey than in the IPUMS sample, though only slightly. In both countries, the share of women in large households of 13 or more persons in much lower in DHS than in the IPUMS sample. The pattern is different in Senegal and Tanzania, where average household size is larger in DHS than in the IPUMS sample. In Senegal, the share of women living in households of nine or fewer persons is slightly lower in the DHS survey. For households of 10 to 12 persons, the share is about the same in both data sources. For large households of 13 or more persons, the share is higher in DHS than in the IPUMS sample. In Tanzania, the share of women in small households of one to three persons is lower in DHS than in the IPUMS sample. The share of women in households of four to six persons is slightly higher in DHS than in the IPUMS sample. The share of women in households of seven to nine people is about the same across data sources, and the share of women in households of 10 or more persons is slightly higher in the DHS survey.

Figure 2.2 Select indicators for women aged 15 to 49 by data source, Benin, Guinea, Senegal and Tanzania



Source Authors' elaboration based on data from IPUMS International and IPUMS DHS

Figure 2.2 cont. Select indicators for women aged 15 to 49 by data source, Benin, Guinea, Senegal and Tanzania



Source Authors' elaboration based on data from IPUMS International and IPUMS DHS

In the final stage of the analysis, individual-level microdata for women aged 15 to 49 years from DHS surveys and IPUMS census samples are pooled by country for use in logistic regression models that estimate the odds that an individual observation derives from DHS. Table 2.2 presents the odds ratio results of logistic regression models predicting the likelihood that a given individual-level observation in each country pooled dataset has DHS rather than IPUMS as its source. Model 1 includes age and number of children ever born as the only independent variables and confirms generally higher odds that an individual observation comes from DHS rather than an IPUMS sample for women who have children compared to women who have no children across all four countries. These results confirm that having children increase the odds that an individual observation in the pooled country datasets comes from DHS, and suggests that childless women are less often captured in DHS than in the census.

In Benin, Guinea, and Tanzania, the odds of DHS as the data source are higher among women who have had children for all but the highest-parity women, although the statistical significance varies by number of children born. In Senegal, the odds of DHS as the data source are higher for women who have had children compared to those who have not across all parities. In Benin, the odds of DHS as the data source are lower for women with 9 or more births compared to those with no births; in Guinea and Tanzania the odds of DHS as the data source are lower for women with 10 or more births compared to women with no births. In Benin and Guinea, this is consistent with the finding that the share of women in very large households is higher in the IPUMS sample than in the DHS survey. In Tanzania, large households of 10 or more persons are less common than in the other countries included in the analysis; a small sample of women with 10 or more births may explain why the trend in odds differs for these women. In contrast to the pattern in Benin, Guinea, and Tanzania, the odds of DHS as the data source are largest for the highest-parity women in Senegal relative to women with no children ever born. In Senegal, where large households are more common, the odds of DHS as the data source are more than twice as high for women with 8 or more births compared to women with no births.

The relationship between parity and data source observed in Model 1 remains robust in Benin and Guinea when we control for individual-level demographic and socioeconomic characteristics in Model 2, suggesting that the two data sources are capturing women with fundamentally different fertility behavior. When marital status, level of education, residence in an urban or rural area, and household size are taken into consideration, the odds of an individual observation deriving from DHS are higher for women who have children compared to those who have not had children for all but those with the highest-parity (10 or more children ever born) in Benin and for all women with children in Guinea. In Benin, the odds of DHS as the data source are 1.5 to 1.7 times higher for women with one to six births compared to women with no births, and 1.1 to 1.4 higher for women with seven to nine births. The odds of DHS as the data source for the highest-parity women (10 or more children every born) are lower relative to women with no children. The effect of parity is stronger in Guinea. The odds of DHS as the data source are 1.8 times higher for women with 1 to 4 births and more than twice as high for women with 5 to 9 births relative to women with no births in Guinea. In Guinea, the odds of DHS as the data source remain 1.3 times higher relative to women with no children ever born for women with 10 or more children ever born. Even when determinants of fertility are controlled for, higher odds of DHS as the data source for women with children point to a possible bias towards women with children in DHS surveys.

In Senegal, the effect is less consistent across parities than in Benin and Guinea when we control for other individual-level characteristics. The odds of DHS as the data source are at least 1.6 times for women who have had 8 or more births compared to women with no births. The effect of parity is lower in magnitude for women with 6 or 7 births, with odds of DHS as the data source that are 1.1 times and 1.3 times higher, respectively, compared to women with no births. Among women with 5 or fewer births, the effect of parity is negligible and statistically less significant, with odds of DHS as the data source that are only slightly higher or slightly lower compared to women with zero births. Consistent with the gaps in share of women living in large households in Senegal presented in figure 2.2, high-parity women have higher odds of being captured in the DHS sample than in the IPUMS sample.

The situation in Tanzania is unique. When we control for marital status, educational attainment, urban-rural status, and household size, the odds of DHS as the data source are lower for women with children compared to women without children. This suggests that differences in other female characteristics rather than number of children born drive differences between data sources. The odds of DHS as the data source are two to four times higher for women who are currently or previously married compared to single women. The odds of DHS as the data source are slightly lower for women who have completed primary education and much lower for women who have completed secondary or higher education. The odds of DHS as the data source are also lower for women living in urban areas compared to those living in rural areas.

Individual-level logistic regression results support the findings from the first phases of the analysis based on aggregate-level data which suggest that different types of households are captured in DHS surveys and censuses. Even in Benin and Guinea where average household size is generally smaller in DHS surveys than in IPUMS census samples, women with children have higher odds of DHS as the data source compared to women without children across parities with the exception of the highest parity women in Benin. In Senegal, women with children have higher odds of DHS as the data source than women without children, though this is only the case for higher-parity women when we control for other individual characteristics. Given the high average household size in Senegal, it may be less common to find childless households. In Tanzania, women with children have lower odds of DHS as the data source compared to women without children when we control for other characteristics. Results suggest that differences between DHS surveys and census samples are linked to marital status, educational attainment, and place of residence rather than number of children born.

Table 2.2 Results of logistic regression of likelihood that DHS is data source of individual observation (women aged 15 to 49 years)

	Odds Ratio (SE)			
	Benin		Guinea	
Variable	*Model 1*	*Model 2*	*Model 1*	*Model 2*
Children ever born				
0 (reference)	1	1	1	1
1	1.39 (.04) ***	1.48 (.05) ***	1.56 (.06) ***	1.82 (.07) ***
2	1.43 (.04) ***	1.56 (.06) ***	1.52 (.06) ***	1.87 (.08) ***
3	1.45 (.05) ***	1.66 (.06) ***	1.40 (.06) ***	1.79 (.09) ***
4	1.38 (.05) ***	1.64 (.07) ***	1.42 (.06) ***	1.87 (.09) ***
5	1.28 (.05) ***	1.59 (.07) ***	1.58 (.07) ***	2.11 (.11) ***
6	1.28 (.05) ***	1.63 (.08) ***	1.56 (.08) ***	2.13 (.12) ***
7	1.07 (.05)	1.38 (.08) ***	1.51 (.09) ***	2.08 (.13) ***
8	1.01 (.06)	1.32 (.08) ***	1.61 (.10) ***	2.26 (.16) ***
9	0.82 (.07) *	1.08 (.09)	1.51 (.12) ***	2.14 (.18) ***
10+	0.64 (.05) ***	0.84 (.07) *	0.95 (.09)	1.34 (.12) **
Age	1.01 (.01)	1.02 (.01) ***	0.95 (.01) ***	0.98 (.01) **
Marital status				
Single (reference)		1		1
Married		0.69 (.02) ***		0.58 (.02) ***
Previously married		0.81 (.04) ***		0.82 (.05) **
Educational attainment				
Less than primary (reference)		1		1
Primary		0.95 (.02) **		1.05 (.03)
Secondary or higher		0.63 (.03) ***		0.92 (.05)
Urban rural status				
Rural (reference)		1		1
Urban		0.92 (.02) ***		0.97 (.02)
Household size		0.91 (0) ***		0.92 (0) ***

***p < 0.001 | ** p < 0.01 | * p < 0.05

Source Authors' elaboration based data from IPUMS International and IPUMS DHS

Table 2.2 cont. Results of logistic regression of likelihood that DHS is data source of individual observation (women aged 15 to 49 years)

	Odds Ratio (SE)			
	Senegal		Tanzania	
Variable	*Model 1*	*Model 2*	*Model 1*	*Model 2*
Children ever born				
0 (reference)	1	1	1	1
1	1.18 (.04) ***	1.04 (.05)	1.03 (.04)	0.77 (.03) ***
2	1.12 (.05) **	0.97 (.05)	1.11 (.04) ***	0.69 (.03) ***
3	1.20 (.05) ***	1.00 (.05)	1.19 (.05) ***	0.64 (.03) ***
4	1.15 (.06) **	0.94 (.05)	1.16 (.05) ***	0.56 (.03) ***
5	1.12 (.06) **	0.90 (.05)	1.20 (.06) ***	0.53 (.03) ***
6	1.41 (.08) ***	1.11 (.07)	1.10 (.06)	0.46 (.03) ***
7	1.61 (.10) ***	1.26 (.09) ***	1.28 (.08) ***	0.50 (.03) ***
8	2.10 (.15) ***	1.62 (.13) ***	1.19 (.08) **	0.45 (.03) ***
9	2.66 (.23) ***	2.05 (.18) ***	1.31 (.10) ***	0.48 (.04) ***
10+	3.07 (.26) ***	2.36 (.21) ***	0.86 (.07) *	0.31 (.02) ***
Age	0.99 (.01)	0.99 (.01)	0.98 (.01) *	1.01 (.01)
Marital status				
Single (reference)		1		1
Married		1.10 (.04) *		1.99 (.06) ***
Previously married		1.61 (.10) ***		3.91 (.16) ***
Educational attainment				
Less than primary (reference)		1		1
Primary		0.79 (.02) ***		0.89 (.02) ***
Secondary or higher		0.57 (.04) ***		0.06 (.01) ***
Urban rural status				
Rural (reference)		1		1
Urban		1.02 (.03)		0.65 (.02) ***
Household size		1.06 (.01) ***		1.10 (0) ***

***p < 0.001 | ** p < 0.01 | * p < 0.05

Source Authors' elaboration based data from IPUMS International and IPUMS DHS

DISCUSSION

The availability of harmonized microdata from censuses and international survey programs has led to an increase in large-scale, cross-national demographic studies that rely on data from multiple sources. Studying the rapidly changing population dynamics of medium- and high-fertility countries requires frequent monitoring over time, creating a need to combine observations from all available sources. While studies that pool data from censuses and surveys are becoming more common, there is little evidence available regarding the coherence of basic demographic indicators and estimates across these data sources. To fill this gap, our analysis explored the comparability of estimates of household size and composition derived from census samples available from IPUMS International and DHS surveys in 22 LMICs. The results indicate systematic differences in the households and women captured across the two sources, with implications for many aspects of demographic research.

The various phases of analysis, based on both aggregate and individual-level data from census and DHS, indicate that differences in the number of children in households drive inconsistencies in household size between DHS surveys and IPUMS census samples.

Using interpolated, aggregated census and DHS survey data available in the CORESIDENCE Database, we have documented differences in average household size between census samples and DHS surveys across 124 country-year observations. We found that for nearly three-quarters of country-year observations, average household size is larger in DHS than in the census sample. The mean absolute difference in household size between IPUMS and DHS observations is 0.31 persons, and the absolute difference is 0.5 persons or higher for 20 percent of observations. Linear regression models that estimate the relationship between average household size and data source when controlling for other household characteristics confirm larger average household size in DHS surveys compared to IPUMS census samples and indicate that differences in the average number of adult women and children across sources are driving discrepancies in household size.

To validate these findings, we used microdata from IPUMS International and IPUMS DHS collected within a two-year period for Benin, Guinea, Senegal and Tanzania to assess the characteristics of women aged 15 to 49 years being captured in each data source. This analysis indicates differences between data sources in fertility and characteristics linked to fertility. In all four countries, a larger share of women living without children—either women who have no children of their own or who do not cohabit with any of their children—are reported in the IPUMS census samples than in the DHS survey. In Benin, Senegal, and Tanzania, the share of single women is higher in the IPUMS sample than in DHS.

The results of logistic regression models estimating that a given individual-level observation in each country pooled dataset has DHS rather than IPUMS as its source confirms that generally, the odds are higher that an individual observation comes from a DHS survey for women who have children compared to women who have no children across all four countries. In Benin, Guinea, and Senegal, the odds of being included in a DHS survey are higher for women with children than for women without children even when controlling for marital status, level of education, urban or rural place of residence, and household size, at least for some parities. Even in Benin and Guinea, where average household size *is* smaller in DHS than in the census sample, the odds of DHS as the data source are higher for women with children than for women without children. In Tanzania, differences in these individual characteristics across data sources drive higher odds of DHS as the source for women with children.

Childless women are more likely to participate in the labor market and live alone or in smaller households. For this reason, they and their households may be more difficult to capture in household surveys. While census enumerators might make multiple attempts to reach the household of a childless woman, a survey interviewer may be quicker to substitute her household for another in the same enumeration area. Similarly, an absent childless woman may be captured in the household roster of a DHS survey and identified as eligible for the individual questionnaire, but the individual questionnaire may never be conducted.

Discrepancies might also be related to inaccurate or outdated sampling frames for certain DHS surveys. Although most DHS survey operations involve a comprehensive household listing exercise as part of the sample selection, the underlying sample frame is often the previous population census, which may have been conducted up to ten years before the survey. This is the case for three of the four countries covered in the individual analysis. In Benin, Senegal, and Tanzania, the sampling frame for the DHS survey was the previous census, conducted ten, eleven, and eight years, respectively, before the survey. For the 2012 DHS in Guinea, the sampling frame was based on mapping work already underway for the 2014 census. The impact of inaccurate sampling frames is not necessarily unique to measures of household size or structure, but smaller households or households without children may be more likely than larger households to move between censuses, impacting their representation in the survey samples.

The results may also point to coverage errors and/or interviewer effects in DHS surveys. DHS surveys are designed to cover a nationally representative sample of households regardless of their composition. Nonetheless, sample size targets are aimed at producing reliable estimates for women aged 15 to 49 and children under five. For this reason, interviewers may deviate from the sample design to favor households with young children. Large gaps between data sources in the share of women with zero children living at home further support this possibility. There may also be a tendency to avoid very large households in some contexts. In Benin and Guinea for example, the share of women living in large households is higher in census samples than in DHS samples. This could point to a "goldilocks" effect of sorts, where interviewers seek households with *some* but not too many children.

These findings partially align with existing assessments of DHS data quality which suggest interviewers and/or respondents may displace or omit births to avoid lengthy child-health questionnaires. If interviewers or respondents displace certain births to inflate the age of the child and avoid the questionnaire for children under age five, overall average household size may still be larger in DHS than in IPUMS samples. Likewise, the displacement of births will not impact the share of childless women captured, which we find to be lower in DHS surveys than in IPUMS samples. On the other hand, the omission of births would contribute to smaller rather than larger average household size in DHS compared to census samples. It may also artificially inflate the share of childless women in DHS samples. Our findings therefore may be consistent with situations of displacement but not omission of births during DHS fieldwork.

At 0.3 persons on average, differences in average household size between IPUMS and DHS are noteworthy but, in our opinion, do not preclude demographic studies that combine these data sources. Many individual-level analyses are unaffected by these discrepancies. Nonetheless, researchers combining data sources to study household size, composition, and change across countries or over time should be aware of potential differences by data source and consider them in the interpretation of their results.

At the same time, the potential underrepresentation of childless women in DHS samples has significant implications for fertility estimates and analyses of the determinants of fertility intentions and outcomes. Childless women are different from women with children in a number of ways, and their underrepresentation may bias results of studies of women of childbearing age on a number of topics including union formation and living arrangements, migration and urbanization, educational attainment, labor market participation, health, intergenerational relations, gender disparities and socio-economic inequalities. Future research in this area could explore differences in indirect measures of fertility across DHS surveys and censuses, providing more precise information on the implications of the findings of our analysis. At the time of writing, the long-term future of the DHS program is uncertain. If the availability of DHS or other survey data become restricted, research based on existing DHS data may become even more important for shedding light on fertility behaviour and household and family structure during a period of significant demographic change in low- and middle-income countries. Our findings should encourage researchers designing studies that pool data across sources to assess the coherence of relevant measures as a part of the design and interpretations of their analyses.

APPENDIX 2

Table A2.1 Difference in average household size between census samples from IPUMS and DHS surveys in number of persons, by country and year (IPUMS less DHS)

Country	Year	Difference in household size (number of persons)	Empirical source
Bangladesh	1993	0.29	DHS (IPUMS)
	1996	0.29	DHS (IPUMS)
	1999	0.50	DHS (IPUMS)
	2001	0.52	Census
	2004	0.50	DHS (IPUMS)
	2007	0.40	DHS (IPUMS)
	2011	0.42	Both (IPUMS)
Benin	1996	0.29	DHS (IPUMS)
	2001	-0.36	DHS (IPUMS)
	2002	-0.36	Census
	2006	-0.50	DHS (IPUMS)
	2011	-0.64	DHS (IPUMS)
	2013	-0.54	Census
Bolivia	1994	0.29	DHS
	1998	0.21	DHS
	2001	0.22	Census
	2003	0.27	DHS
	2008	0.24	DHS
Cambodia	2000	0.30	DHS
	2004	0.20	Census
	2005	0.18	DHS
	2008	0.21	Census
	2010	0.23	DHS
	2013	0.27	Census
Cameroon	1991	0.65	DHS (IPUMS)
	1998	0.46	DHS (IPUMS)
	2004	-0.22	DHS (IPUMS)
	2005	-0.19	Census
Colombia	1990	0.06	DHS
	1993	0.16	Census
	1995	0.19	DHS
	2000	0.20	DHS
	2005	0.39	Both
Dominican Republic	1991	0.22	DHS
	1996	0.13	DHS
	1999	0.21	DHS
	2002	0.06	Both
	2007	0.06	DHS
	2010	0.10	Census
Egypt	1992	0.99	DHS (IPUMS)

Country	Year	Difference in household size (number of persons)	Empirical source
	1995	0.77	DHS (IPUMS)
	1996	0.77	Census
	2000	0.87	DHS (IPUMS)
	2003	0.82	DHS (IPUMS)
	2005	0.77	DHS (IPUMS)
	2006	0.72	Census
Ethiopia	2000	0.33	DHS (IPUMS)
	2005	0.51	DHS (IPUMS)
	2007	0.45	Census
Guinea	1999	-0.24	DHS (IPUMS)
	2005	-0.85	DHS (IPUMS)
	2012	-0.80	DHS (IPUMS)
	2014	-0.86	Census
Haiti	1994	0.41	DHS
	2000	0.06	DHS
	2003	0.02	Census
Indonesia	1991	0.18	DHS
	1994	0.19	DHS
	1995	0.20	Census
	1997	0.20	DHS
	2000	0.41	Census
	2002	0.44	DHS
	2005	0.32	Census
	2007	0.27	DHS
	2010	0.24	Census
Kenya	1993	0.19	DHS (IPUMS)
	1998	-0.09	DHS (IPUMS)
	1999	-0.03	Census
	2003	0.05	DHS (IPUMS)
	2008	-0.12	DHS (IPUMS)
	2009	-0.17	Census
Malawi	1992	0.28	DHS (IPUMS)
	1998	0.17	Census
	2000	0.10	DHS (IPUMS)
	2004	-0.04	DHS (IPUMS)
	2008	0.04	Census
Mali	1995	-0.33	DHS (IPUMS)
	1998	-0.59	Census
	2001	-0.75	DHS (IPUMS)
	2006	-0.44	DHS (IPUMS)
	2009	-0.43	Census
Philippines	1993	0.28	DHS
	1995	0.27	Census
	1998	0.19	DHS

Country	Year	Difference in household size (number of persons)	Empirical source
	2000	0.10	Census
	2003	0.04	DHS
	2008	0.22	DHS
	2010	0.28	Census
Rwanda	1992	0.24	DHS (IPUMS)
	2000	-0.12	DHS (IPUMS)
	2002	-0.08	Census
	2005	0.06	DHS (IPUMS)
	2007	-0.10	DHS (IPUMS)
	2010	0.14	DHS (IPUMS)
	2012	0.13	Census
Senegal	1992	0.00	DHS (IPUMS)
	1997	0.09	DHS (IPUMS)
	2002	-0.21	Census
	2005	-0.11	DHS (IPUMS)
	2010	0.91	DHS (IPUMS)
	2012	0.81	DHS (IPUMS)
	2013	0.72	Census
Tanzania	1991	0.54	DHS (IPUMS)
	1996	0.22	DHS (IPUMS)
	1999	0.47	DHS (IPUMS)
	2002	0.46	Census
	2004	0.39	DHS (IPUMS)
	2007	0.54	DHS (IPUMS)
	2010	0.44	DHS (IPUMS)
	2012	0.41	Census
Turkey	1993	0.03	DHS
	1998	0.13	DHS
	2000	0.10	Census
Uganda	1995	0.11	DHS (IPUMS)
	2000	0.14	DHS (IPUMS)
	2002	0.19	Census
	2006	0.31	DHS (IPUMS)
	2014	-0.06	Census
Zambia	1992	-0.14	DHS (IPUMS)
	1996	-0.06	DHS (IPUMS)
	2000	0.10	Census
	2001	0.07	DHS (IPUMS)
	2007	-0.30	DHS (IPUMS)
	2010	-0.18	Census

Note All census samples accessed from IPUMS International
Source Authors' elaboration based on data from IPUMS International, IPUMS DHS, and DHS Program

CHAPTER 3

NON-FAMILY LIVING ARRANGEMENTS AMONG YOUNG ADULTS IN THE UNITED STATES

INTRODUCTION

Transformations in household composition and living arrangements in the United States have accelerated in recent years. Households are smaller, more people live alone or with unmarried partners, and adult children are prolonging and returning to coresidence with their parents. One of the least studied aspects of these household changes is the rise of non-family living arrangements. According to the 2019 American Community Survey (ACS), nearly 10 percent of individuals aged 20 to 39 live with only non-relatives, yet we know very little about the long-term trends in non-family living arrangements among young adults and the characteristics of individuals living with non-relatives.

To fill this gap, this study undertakes an exploratory analysis of trends in non-family living arrangements among young adults aged 20 to 39 in the United States and the associated individual characteristics. We begin by describing the recent change to patterns of family formation in the United States and how these changes have impacted living arrangements among young adults. We use US census and American Community Survey (ACS) data collected between 1990 and 2019 to document evolving patterns in non-family family living arrangements over time, identifying persons in non-family living arrangements as those who are not related by kinship or partnership to any other person in the household. We examine the demographic and socio-economic characteristics associated with non-family living.

Results indicate that non-family living arrangements among young adults aged 20-39 are increasingly common and not only during the ages associated with third-level education. We find that non-family living is associated with markers of both advantage and disadvantage, and that differences across age groups point to parallel profiles of those living with non-relatives that diverge with age. Trends among younger age groups are linked more closely to socio-economic patterns around family formation. Among older young adults, the demographic and labor force characteristics of the foreign born and constraints of their kin availability drive trends. Our study provides relevant evidence around an increasingly common living arrangement in the United States and also identifies several areas for future research on the diversity of living arrangements among young adults in the United States and the implications of these trends.

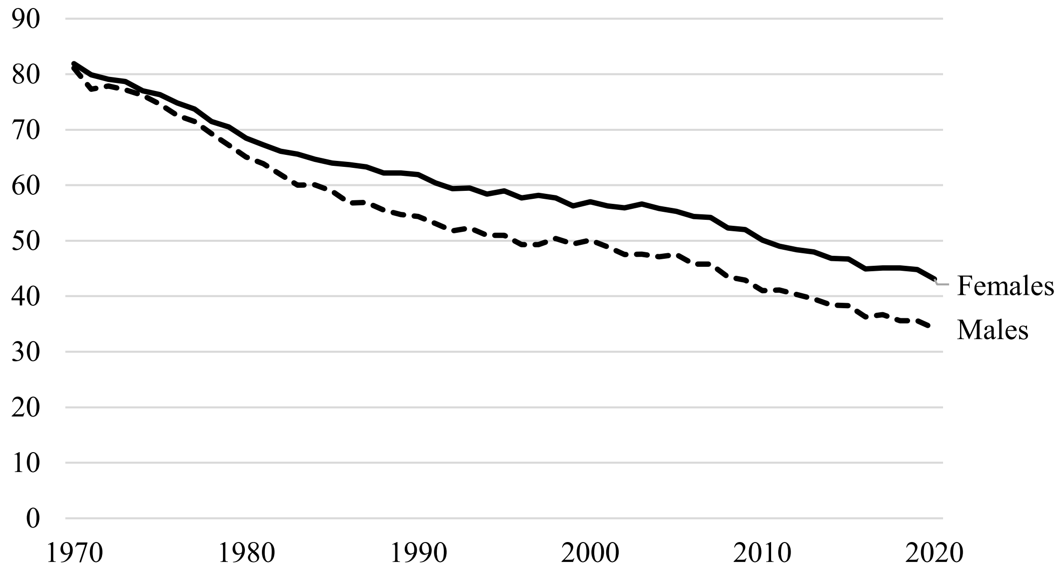
BACKGROUND

SHIFTS IN FAMILY FORMATION PATTERNS

Shifts in family formation patterns in the United States during the last several decades are well-documented in the literature (Cherlin 2010; Kuperberg 2014; Lesthaeghe and Neidert 2006; Manning 2020; Miller 2020; Ruggles 2015; Sassler and Lichter 2020; Smock and Schwartz 2020). While trends vary across regions and population groups, overall, the characteristics of family formation in the United States align with the features of the Second Demographic Transition observed also in Western Europe and other high-income countries: delayed marriage and more cohabitation among couples, later and lower fertility, and increased childbearing outside of marriage (Lesthaeghe and Neidert 2006). The median age at first marriage reached its peak in the United States in 2020, climbing steadily to 30.5 for men and 28.1 for women from 23.2 at 20.8 respectively in 1970 (U.S. Census Bureau 2020a). In 1976, fewer than one-third of women aged 25-29 were childless. By 2018, more than half of women this age were childless (U.S. Census Bureau 2019). Delayed childbearing has contributed to a decline in the average number of children born to women in the United States. While the Total Fertility Rate has been below replacement level (2.1) since 1971, it reached a record low of 1.64 in 2020 (Osterman et al. 2022). Like overall fertility rates, nonmarital fertility rates are declining, but the share of all births to unmarried women has increased from around 10 percent in 1970 to 40 percent in 2020 (Osterman et al. 2022; Ventura and Bachrach, 2000).

These demographic shifts have impacted living arrangements across all age groups but particularly among young adults who represent the age group at which many demographic transitions occur. The transition to adulthood—completing education, beginning full-time employment, leaving the parental home, getting married and having children—occurs later and takes longer now than it did in the period after World War II (Furstenberg 2010). Delayed and declining marriage and childbearing means that a larger proportion of young adults do not have spouses and/or children with whom to live.

Figure 3.1 Share of Males and Females Aged 25-34 Living With a Spouse, With or Without Children 1970-2020 (%)



Source Authors' elaboration based on U.S. Census Bureau, Current Population Survey, Annual Social and Economic Supplement, 1967 to present

Consequently, the share of young adults living in married-couple family households has sharply declined in recent decades (see Figure 3.1). In 1970, more than 80 percent of individuals aged 25 to 34 lived with a spouse (with or without a child). By 1995, this had fallen to 55 percent. In 2020, 39 percent of young adults in this age group lived with a spouse (U.S. Census Bureau 2020b). This trend reflects changes in the household composition across all age groups and provides support for earlier observations of a shortening of the portion of the life cycle during which individuals live with family (Kobrin 1976) and a growing disconnection between family and households in the United States (Cherlin 2010). In 1970, more than 80 percent of households in the United States contained families. In 2020, 65 percent of households contained families (U.S. Census Bureau 2020c).

ALTERNATIVES TO MARITAL LIVING ARRANGEMENTS

The alternative living arrangements available to young adults are finite: if one does not live with a spouse or child, one must live with an unmarried partner, with other family members, alone, or with non-family members. Among the alternative living arrangements available to young adults, most research has focused on cohabitation among unmarried couples (Hemez and Manning 2017; Manning 2020; Sassler and Miller 2017; Sassler et al. 2018; Sassler and Lichter 2020; Di Giulio et al. 2019) and, to lesser extent, on the rise in parent-adult child coresidence (Bianchi et al. 2007; Esteve and Reher 2021; Goldscheider and Goldscheider 2007; Kahn et al. 2013; Keene 2010; Lei and South, 2016; Mazurik et al. 2020; Sandberg-Thoma et al. 2015). This attention reflects the considerable increase in these living arrangements during the last several decades. In 1975, only one percent of adults aged 25 to 34 lived with an unmarried partner. In 2020, nearly 15 percent of young adults lived with an unmarried partner (U.S. Census Bureau 2020b). In 2017, 62 percent of women aged 19-44 had ever cohabited compared to one-third in 1987 (Manning 2020). Parent-adult child coresidence has also increased rapidly, especially in recent years and among males. Between 1980 and 2020, the share of males aged 25 to 34 living with a parent nearly doubled from 11 percent to 21 percent (U.S. Census Bureau 2020b). This is now the most common living arrangement among unmarried young adult males. Living alone has also become more common. The share of young adults living alone has more than doubled since 1970 (U.S. Census Bureau 2020b). In 2020, 12 percent of men and 9 percent of women aged 25 to 34 lived alone (U.S Census Bureau 2020b).

Changes in living arrangements and family formation have been linked to attitudinal and economic changes. Cohabitation among unmarried adults has become more accepted in American society, and cohabitation more often than marriage is now a primary marker of emerging adulthood for young people (Furstenberg 2015; Manning and Cohen 2015; Sassler and Litcher 2020). The timing and sequence of the transition to adulthood differs by socioeconomic status (Furstenberg 2008). Socioeconomic differences are also driving trends in cohabitation. Women from advantaged backgrounds are more likely to marry and less likely to cohabit, and college-educated women are more likely to transition from cohabitation to marriage than less-educated women (Manning et al. 2014; Sassler et al. 2018; Smock and Schwartz 2020). Economic stability has become a prerequisite to marriage. Many young people expect to reach a certain level of job security and income before they marry (Furstenberg 2015; Ruggles 2015). In the last 40 years, real wages among young men and women have declined in the United States (Ruggles 2015). These economic barriers contribute to delayed marriage and increased cohabitation and provide support for the diverging destinies framework which highlights differences in demographic change across socioeconomic groups (McLanahan 2004; Smock and Schwartz 2020).

Economic conditions are also driving the increase in parent-child coresidence in the last several decades, a trend that has accelerated rapidly since the Great Recession. The share of young men aged 25 to 34 living with a parent increased from 15 to 21 percent between 2010 and 2020, and the economic downturn associated with the Covid-19 pandemic has likely contributed to further increases since 2021 2021. Challenging the historical understanding of intergenerational coresidence, the arrangement has become associated with economic need among the younger not the older generation (Kahn et al. 2013; Matsudaira 2016; Ruggles 2007; Sironi and Furstenberg 2012). In 1960, adult children with more resources were more likely to live with parents. In 2010, adult children with fewer resources were more likely to live with parents (Kahn et al. 2013). Declining wages, rising housing costs and student debt have contributed to a deteriorating economic position for young people in the United States which has driven them back to their parental homes (Houle and Warner, 2017; Matsudaira 2016; Ruggles 2015). As is the case with cohabitation among unmarried partners, there are important socioeconomic differences. Young adults who are non-white or have low education are more likely to live with a parent (Kahn et al. 2013). This mirrors population-level trends—intergenerational and extended-family households are more common among non-white and low-SES individuals of all ages in the United States (Amorim et al. 2017; Cross 2018; Harvey and Dunifon 2023; Kamo 2000; Pilkauskas 2012; Pilkauskas and Cross 2018; Pilkauskas et al. 2020; Swartz 2009; Reyes et al. 2020).

NON-FAMILY BASED LIVING ARRANGEMENTS

There is comparatively less research on another increasingly common living arrangement among young adults—non-family living. In 2019, nearly 10 percent of individuals aged 18 to 24 and 7 percent of individuals aged 25 to 34 lived with non-relatives only (U.S. Census Bureau 2020b; 2020e). The limited literature that has explored trends in non-family living in the United States has (Christian 1989), by necessity, included cohabiting unmarried partners in this category of living arrangement: before 1990, the U.S. Census did not distinguish unmarried partners from other non-relatives in relationship data. Now that it is possible with available data, it is important to examine coresidence with non-relatives separately from cohabitation.

The existing evidence tells us little about how cultural and economic factors have influenced the rise in non-family living arrangements. One may think that in an individualistic society like the United States, living independently—that is, alone or with a spouse or partner—is preferred to living in an intergenerational or extended-family household, living arrangements that are generally associated with economic disadvantage. Where exactly non-family living arrangements fall on the spectrums of independence and affluence is not clear. Among older adults, non-family living may be a last resort for the economically disadvantaged. However young adults living with non-relatives could represent a relatively privileged segment of the population. Well-educated young adults are more likely to be unmarried and childless, giving them the option to live with non-relatives (Furstenberg 2008). Living away from the parental home before marriage and/or childbearing is an increasingly important rite of passage (Rosenfeld 2007), and living with non-relatives might represent a lifestyle choice among young adults which provides social as well as financial benefits (Heath and Kenyon 2001; Stone et al. 2011). Some recent descriptive evidence on non-family households is available from the Census Bureau, but we know little about the individual characteristics associated with the arrangement.

One way to understand whether non-family living is a response to economic need or reflects a preference regardless of socioeconomic status is to examine the profiles of individuals in this arrangement. Among the several factors that might be driving the trends in non-family living arrangements among young adults, three stand out as having potentially strong and direct effects: education, migration, and income.

The share of the population pursuing university and post-graduate study has increased in the United States with the potential to impact living arrangements in several ways. Many college students live in university collective housing or share private dwellings with classmates. Students are also likely to delay family formation until they have completed their studies. An increase in the number and age of students could therefore prolong periods of university-related non-family living arrangements. The university educated are also more likely to work in specialized professions that require relocation and to pursue professions concentrated in large urban areas with high rents. Indeed, in their study of living arrangements among young adults in the United Kingdom, Stone and colleagues (2011) find that living with non-relatives is more common among students or individuals with a previous experience of higher education. We therefore take into account the influence of educational expansion by restricting our analysis to non-family living arrangements in private households (excluding group quarters such as university dormitories) and including measures of educational attainment and school attendance in our analysis.

Migrants—both international and internal—have historically been more likely to live in complex and non-family households (Van Hook et al. 2007). The availability of kin is restricted for migrants for a number of reasons: young migrants are more likely to be single and childless and migrants of all ages are more likely to live apart from partners, children, and other family members. Furthermore, migrants are more likely to work in low-wage jobs that do not afford independent living especially in large urban areas with high housing costs. Stone and colleagues (2011) find that the increase in non-family living among young men in the United Kingdom is largely explained by immigration-driven changes in the composition of the population.

The need to share housing with non-family members is not unique to migrants. In many cities in the United States housing costs prohibit independent living. Journalistic evidence suggests it is becoming increasingly common for cohabiting or married couples—even those with children—to live with roommates to help defray high rents (Herron 2019; Moylan 2019; Pappu 2016). Our analysis explores living arrangements such as these from the perspective of the roommate. In line with the literature on other forms of non-independent living arrangements, we expect non-family living to be less common among those with higher incomes. We include measures of migration, income, and urban residence in our analysis.

To complement the literature on other living arrangements among young adults, we address several original questions: What share of individuals live in non-family arrangements across age groups and has this changed over time? Are trends driven by educational expansion and migration? Do non-family living arrangements reflect preferences or necessity? Are they associated with affluence or disadvantage? To answer these questions, we examine trends in the share of young adults living with non-relatives over time and determine which individual characteristics are associated with living in non-family households. Given our interest in individual demographic transitions, we focus our investigation on the period of early adulthood between the ages of 25 and 29 during which most individuals begin to establish independent households, but we present results for other age groups as well.

DATA AND METHODS

Our interest is in shared living arrangements among individuals without family or kin-like ties. We consider non-married partners to have a kin-like relationship and exclude them from our definition of non-family living. The unitary model of household utility proposed by Gary

Becker (1981) has been challenged for family households, and this model is even less applicable to households of non-relatives where the pooling of resources and uniform preferences are less likely. For this reason, we study individuals rather than entire households. We identify individuals living in non-family arrangements as those persons living in multi-person private households who are not related to anyone in the household, without distinguishing whether the other members are related or not.

This study pools U.S. Census microdata for 1990, 2000, and American Community Survey (ACS) microdata for the years 2010 and 2019. These data provide the longest available view of change in the prevalence of non-family living in the United States. The 1990 census is the earliest to distinguish non-married partners from other non-relatives, and the 2019 ACS is the most recent national survey that is sufficiently comparable to earlier data. Microdata for the 2020 ACS were available at the time of writing, but due to the effects of the COVID-19 pandemic on data collection and data quality, the Census Bureau released the 2020 1-year ACS PUMS file with experimental weights and cautioned that the data are not comparable to other ACS sample years. Furthermore, the COVID-19 pandemic disrupted the living situations of many individuals—especially college students and young adults. Data from 2020 are therefore not comparable to previous years or useful for studying long-term trends. These sample years provide roughly equally spaced observations across four decades. The microdata samples were accessed through IPUMS and provide self-weighted, nationally representative 5 percent samples of the 1990 and 2000 censuses, and 1 percent samples of the ACS (Ruggles et al. 2021).

In these microdata samples, individuals are organized into households, allowing for the analysis of household structure and classification of household type. To identify individuals in private non-family living arrangements, we rely on the relationship to household head variable, RELATE, and the IPUMS-constructed family interrelationship variables that identify each individual's co-resident spouse (opposite or same-sex), mother, and father. The RELATE variable indicates if and how each individual living in the household is related to the household head. The interrelationship variables allow us to identify family ties among individuals not related to the head.

IPUMS assigns links between family members in the family interrelationship variables based on the information reported in the RELATE variable as well as age, sex, marital status, and the order in which individuals are listed on the census and survey forms (more information on the

IPUMS family interrelationship variables is available from the IPUMS website and in Ruggles 1995). To illustrate the value of these family interrelationship variables, take the household presented in Table 3.1 as an example. The empirical relationship information collected in the census or ACS is in reference to the household head, and the relationships among persons 2, 3, and 4 are not documented in any variable reflecting data collected directly from respondents in questionnaires. Nonetheless, the age, sex, and marital status information of these individuals as well as their proximate order in the household roster point to family ties. The IPUMS samples use this information to link these individuals and generate "pointer" variables indicating the location in the household of spouse and parents. Using these variables, we can identify more family relationships than based on relationship-to-head alone.

Table 3.1 Family relationships in IPUMS

Person	Relationship	Age	Sex	Marital status	Spouse	Mother	Father
1	Head	64	F	Widow	0	0	0
2	Non-relative	31	M	Married	3	0	0
3	Non-relative	29	F	Married	2	0	0
4	Non-relative	4	F	Single	0	3	2

Source Authors' elaboration

We use a process of elimination to identify individuals living with non-relatives. We first exclude all individuals who are related to the head of the household. We make no distinctions between primary kin and extended family. Any individual related to the head by blood or marriage is considered to live with family. The heads of these households are also excluded. Foster children are distinguishable from other non-relatives beginning in 2000. In our analysis, foster children are considered non-relatives, but they represent a very small percentage of individuals in the age groups of focus. Next, we exclude individuals for whom co-resident spouses, mothers, fathers, or children have been identified based on the interrelationship pointer variables. The individuals that remain fall into two categories 1) individuals who are not related to the head of household and no spouse, mother, father, or child has been identified in the household and 2) heads of households that contain only non-relatives of the head (person 1 in Table 3.1, for example). We classify these individuals as living in non-family arrangements. The pooled sample consists of 32,002,596 individuals and 12,348,163 households. Individuals living in households where no one is related—"roommate" households—represent 71 percent of the sample, and individuals not related to anyone in a household with related persons—persons "hosted" by a family household—represent 29 percent of the sample. Results presented in the following section are consistent whether or not individuals in the second group are included.

We do not capture every family relationship with this approach. We do not identify persons unrelated to the head but related to one another in ways other than parent-child or spouse. For example, if two siblings or cousins live together in a household with other individuals and neither of the pair of siblings or cousins is related to the household head it is not possible to identify this family relationship. These individuals are not excluded in our process of elimination and are counted among those living in non-family arrangements. This could lead to a slight overestimation of the size of the population living with non-relatives only.

The analysis consists of two parts. The first phase documents trends in non-family living arrangements in the United States by age group across time. Anticipating both age and period effects, we document age patterns in non-family family living arrangements and how these patterns have changed across time. The second phase of the analysis relies on logistic regression to estimate the likelihood of living with non-family when we control for individual-level demographic and socio-economic characteristics.

We focus our analysis on the period of early adulthood between the ages of 25 and 29 during which most individuals begin to establish independent households. Results are also presented for the 20-24, 30-34, and 35-39 age groups.

We run two logistic regression models each for males and females to examine changes over time with and without adjustments for other sociodemographic characteristics. The dependent variable in the models is an indicator of whether the individual lives with non-family only. Model 1 considers period effects only with census or survey year as the only independent variable. In Model 2, in addition to census or survey year, we include other individual characteristics as potential correlates of an individual's likelihood of being in a non-family living arrangement. Marital status includes three categories: (1) never married, (2) married, and (3) separated, divorced, or widowed. We combine IPUMS race and Hispanic origin variables into a single measure with five categories: (1) non-Hispanic White, (2) non-Hispanic Black, (3) non-Hispanic Asian, (4) non-Hispanic other race, and (5) Hispanic/Latino of any race. Nativity is dichotomized based on the individual's place of birth: (1) native (born in the United States, excluding outlying areas and territories) and (2) foreign-born. School attendance corresponds to the IPUMS variable which indicates for individuals aged three years and older if he or she was attending nursery school, kindergarten, elementary school, or any schooling leading toward a high school diploma or college degree. Educational attainment categorizes individuals based on highest year of school or degree *completed* at the time of the survey or census: (1) primary or less: grade 11 or below completed, (2) secondary: grade 12 completed or up to three years of college completed, (3) four years or more of college completed. Employment status indicates whether the individual was (1) employed, (2) unemployed, or (3) not in the labor force during the week before the census or ACS was completed. We categorize whether the individual's household was located in a metropolitan area according to the IPUMS metropolitan status variable, which uses the U.S Census Bureau definition: a region consisting of a large urban core together with surrounding communities that have a high degree of economic and social integration with the urban core. Poverty status indicates if the individual had a family total income for the previous year (1) below the official poverty threshold, (2) between 100 and 199 percent of the poverty threshold, or (3) 200 percent of poverty threshold or above. Poverty status is determined at the family—not the household—level. Therefore, for adults living as unrelated individuals, poverty status reflects individual income.

We have not included interaction terms in our model. We explored interactions between independent variables and did not observe strong categorical differences in the effect of independent variables on one another.

The variables included in the analysis are highly comparable across the samples pooled for analysis. Basic demographic variables such as age, sex, and marital status are completely comparable for our purposes. Practices around the collection of race and educational attainment information have been generally consistent across censuses and ACS in the US since 1990. IPUMS-constructed income and poverty variables provide cross-temporally comparable measures that account for inflation and changes to poverty thresholds. Likewise, the IPUMS-constructed variable that identifies households in metropolitan areas applies consistent metropolitan area definitions across all samples included in the analysis.

RESULTS

We present results for the analysis of non-family living arrangements in the United States by age group and across time separately for males and females. Table 3.2 reports the percent living with non-family for men and women at ages 20-24, 25-29, 30-34, and 35-39 by census or survey year. The share of individuals living with non-family has increased over time at every age. Change over time has been monotonic at most but not all ages. The share of individuals in non-family living arrangements is highest in 2019 for all age groups except for men aged 20-24. Among men aged 20-24, the share living with non-family is highest in 2010. Nonetheless, the share of men aged 20 to 24 living with non-family is higher in 2019 than in 1990. We do not observe this pattern among men at other ages or among women, for whom there has been a steady increase between the earliest observation and the most recent observation at all ages. Men are more likely than women to live with non-family at all ages, and the difference in the share of living with non-family between the earliest and most recent observations is greater for men than for women across the 25-29, 30-34, and 35-39 age groups. The difference between the earliest and most recent observations is greater for women than men at age 20-24. The largest increases over time are observed at ages 25-29 for men and ages 20-24 for women. Between 1990 and 2019, the share of men living with non-family at ages 25-29 increased by 4.3 percentage points. For women aged 20-24, the share living with non-family increased by 4.2 percentage points between 1990 and 2019. The change over time is less pronounced at ages 30-34 and 35-39, and remains slightly larger for men than for women.

Table 3.2 Percentage of males and females aged 20 to 39 living in non-family households by 5-year age group and census or survey year

Year	Males				Females			
	20-24	25-29	30-34	35-39	20-24	25-29	30-34	35-39
1990	15	10	5.6	3.9	12.1	6.3	3.4	2.4
2000	18.3	12.2	7	4.6	14	7	3.6	2.7
2010	18.4	13.7	8	5.6	15.6	8.4	4.2	2.8
2019	17.0	14.3	9.0	5.7	16.3	10.1	5.0	3.4

Source Authors' elaboration based on data from IPUMS USA

Table 3.3 provides sample characteristics and presents the share of men and women in non-family living arrangements by covariates included in the logistic regression models and census or survey year for the 25-29 age group (sample characteristics and bivariate relationships for other age groups are available in Appendix tables A3.1 and A3.2). Considerable changes in the composition of the sample population over time are observed. The share of 25-to-29-year-olds completing university has increased, with a larger increase for females. School attendance increased between 1990 and 2010 and decreased slightly between 2010 and 2019. Employment has decreased among men but increased among women. The relative size of the foreign-born population increased significantly between 1990 and 2010 and has decreased since. The share of the population that is white has decreased steadily. Among the covariates considered, the largest compositional change has been around the share of individuals who are single at ages 25-29. Nearly three-quarters of men and two-thirds of women were single at this age in 2019, compared to less than half of men and less than one-third of women in 1990. The share of men and women living in urban areas has increased, but only slightly. Poverty was highest among both men and women in 2010 and has decreased since. Across all years, women are more likely than men to experience poverty at this age.

The bivariate results presented in Table 3.3 tell a somewhat paradoxical story about the relationship between socioeconomic status (SES) and the likelihood of living with non-family only. Higher absolute levels in the proportion of people in non-family living arrangements and larger changes over time are observed for specific subgroups, but trends across characteristics are not consistent with typical conceptualizations of advantage and disadvantage. Compared to the total population ages 25-29 living with non-family (presented at the bottom of Table 3.3), men and women with a university degree are more likely to live with non-family, implying the living arrangement may be associated with higher SES. Among men, the change over time for those with a university degree is larger than for the total population in this age group. The change over time is about the same among university-educated women and the total population. Non-family living arrangements are associated with university students, but across all years individuals attending school are less likely to live with non-family at this age (females only slightly so) compared to the total population.

The period gradient among those who were employed, were non-Hispanic white, or were living in urban areas largely mirror trends for the total population, though the share of individuals living in non-family arrangements in these groups were slightly higher compared to the overall population in all years. The share living with non-family is higher for foreign-born individuals than the total population in all years. In 2010 and 2019, nearly 20 percent of foreign-born men lived with non-relatives only. Incongruous with trends around educational attainment, non-family living is much more common among individuals with incomes below the poverty threshold. More than one quarter of men (27 percent) and 17 percent of women with incomes below the poverty threshold lived in non-family households in 2019. The difference over time is much more pronounced for individuals with incomes below the poverty level than for the total population aged 25-29 living with non-family as well, with the share increasing by nine and seven percentage points between 1990 and 2019 for men and women respectively. The share of single people living in non-family arrangements is higher than for the total population, but no pattern across time is discernible.

Table 3.3 Sample characteristics and share in non-family households by covariate and census or survey year, males and females aged 25-29

	Males		Females	
	% of total sample	% living in non-family household	% of total sample	% living in non-family household
University complete				
1990	22.3	15.1	22.5	11.1
2000	25.6	16.4	29.9	10.9
2010	27.1	18.4	35	12.5
2019	32.5	19.9	39.9	13.9
Attending school				
1990	13.7	9.1	13.6	5.7
2000	13.6	11.4	15.1	6.3
2010	15.4	13.1	19.5	7.9
2019	12.7	13.6	16.1	9.5
Employed				
1990	87	9.9	69.9	7.5
2000	83.2	12.1	69.7	7.8
2010	78.6	14.3	69.7	9.5
2019	84.7	14.8	77.3	11
Foreign-born				
1990	12.9	14.9	11.4	7.9
2000	20.7	16.6	18.5	8.2
2010	19.8	19.5	18.7	8.9
2019	15.2	19.1	15.3	11.3
Non-Hispanic white				
1990	74	9.8	72.8	6.5
2000	63	11.6	62.4	7.4
2010	59.6	13.4	58.1	9.2
2019	55.2	14.8	54	11.1
Single				
1990	44.2	19	31.2	16
2000	49.7	20.7	38.2	15.1
2010	63.8	19.4	51.8	14.3
2019	71.8	18.6	61.5	15.1
Living in metropolitan area				
1990	79.6	11.6	79.3	7.3
2000	80	13.4	80.1	7.8
2010	81	14.8	81.1	9.3
2019	83.6	15.3	83.6	11
Income below poverty threshold				
1990	10	17.9	14.2	9.7
2000	9.9	22.4	14.6	11.2
2010	13.5	23.2	19.7	12.1
2019	8.9	26.8	14.6	16.8
Total				
1990		10		6.3
2000		12.2		7
2010		13.7		8.4
2019		14.3		10.1

Source Authors' elaboration based on data from IPUMS USA

Table 3.4 presents the odds ratio results of logistic regression models predicting the likelihood of non-family living for men and women aged 25 to 29. Model 1 includes census or survey year as the only independent and confirms higher odds of non-family living in recent years compared to earlier years for both men and women. The odds of living with non-family are 1.5 times higher among males in 2019—compared to the earliest observation in 1990. Among females, the odds of living with non-family are 1.7 times higher in 2019 compared to 1990.

When we control for individual-level demographic and socio-economic characteristics in Model 2, the increases over time disappear, suggesting that the increase in non-family living arrangements among young adults is largely explained by changes in the composition and characteristics of the population over time. Consistent with the bivariate analysis, level of education is strongly positively associated with non-family living. The odds of living with non-family are nearly twice as high for men who have completed four or more years of college compared to those who have completed less than secondary education. Among women, those with a university degree have odds of living in a non-family arrangement that are 2.2 times higher than for their less educated peers. Attending school increases slightly the odds of non-family living for both men and women. Working men and women have higher odds of living in non-family arrangements than the unemployed or those outside the labor force. The odds of living with non-family are 50 percent higher among employed men than among unemployed men. Among women, the odds of living with non-family are 40 percent higher among the employed compared to the unemployed.

Non-family living arrangements are more common among those born outside the US than the native-born. The odds of living with non-family only are 1.8 times higher for foreign-born men than for native-born men, and the odds for foreign-born women are 1.5 times higher than for native-born women. There are significant differences by race. The odds of non-family living are higher among white individuals than among other racial and ethnic groups. The odds of non-family living are lowest among the black and the Hispanic populations. The odds of living with non-family are 60 percent as high among non-Hispanic black men and 80 percent as high among Hispanic men compared to white men. The odds are 40 percent as high for non-Hispanic black women and 60 percent as high for Hispanic women compared to white women. Not surprisingly, the odds of non-family living are vastly lower among married individuals than among individuals who have never been married. Divorced, separated, or widowed men have odds of living with non-family only that are 10 percent lower than never married men. The odds among previously married women are 40 percent lower than among never married women.

Non-family living is an urban phenomenon. Living in a metropolitan area significantly increases the odds of non-family living for both men (odds ratio = 1.65) and women (odds ratio = 1.80). Income is strongly and negatively associated with non-family living for both men and women. Those with personal income levels below the poverty threshold have odds of living with non-family that are 60 (females) to 70 (males) percent higher than those with incomes that are more than twice the poverty threshold.

We also ran Models 1 and 2 for the 20-24, 30-34, and 35-39 age groups (see Appendix tables A3.3-A3.5). The patterns observed for the 25-29 age group are largely mirrored in the other age groups with some notable exceptions. Among men and women aged 20 to 24, education and income levels are the strongest predictors of living with non-family, and the relative effect of nativity status is lower compared to other control variables. Among the older age groups, the effect of education is smaller relative to other control variables. Marital status, nativity status, and income level are the strongest predictors of non-family living among those aged 30 to 34 and 35 to 39. For these age groups, level of education has less importance relative to most other control variables. At ages 35-39, education is negatively associated with non-family living for men. The odds of living with non-family are slightly higher among Non-Hispanic Asian women than among white women at ages 30-34 and 35-39.

Table 3.4 Results of logistic regression of likelihood to live in a non-family household for males and females aged 25-29

	Odds Ratio (SE)			
	Males		Females	
Variable	*Model 1*	*Model 2*	*Model 1*	*Model 2*
Year				
1990 (reference)	1	1	1	1
2000	1.243 (.002)	1.158 (.002)	1.117 (.002)	0.924 (.002)
2010	1.427 (.002)	1.049 (.002)	1.372 (.002)	0.847 (.002)
2019	1.495 (.002)	1.047 (.002)	1.687 (.003)	0.934 (.002)
Educational attainment				
Primary or less (reference)		1		1
Secondary		1.095 (.002)		1.109 (.003)
University		1.890 (.003)		2.198 (.006)
School attendance		1.164 (.001)		1.126 (.002)
Employment status				
Employed (reference)		1		1
Unemployed		0.493 (.001)		0.598 (.002)
Inactive		0.538 (.001)		0.632 (.001)
Foreign-born		1.755 (.002)		1.509(.003)
Race and ethnicity				
White (reference)		1		1
Non-Hispanic Black		0.561 (.001)		0.364 (.001)
Non-Hispanic Asian		0.861 (.002)		0.964 (.003)
Non-Hispanic other		0.779 (.002)		0.765 (.003)
Hispanic any race		0.764 (.001)		0.629 (.001)
Marital status				
Single (reference)		1		1
Married		0.064 (.000)		0.046 (.000)
Divorced, separated, widowed		0.869 (.001)		0.583 (.001)
Residence in metropolitan area		1.647 (.002)		1.803 (.004)
Poverty status				
Below threshold (reference)		1		1
Income 100-200% of threshold		0.578 (.001)		0.655 (.001)
Income > 200% of threshold		0.296 (.000)		0.383 (.001)

p < 0.001 for all variables
Source Authors' elaboration based on data from IPUMS USA

DISCUSSION

This study has investigated trends in non-family living arrangements among young adults in the United States and sheds light on the individual characteristics associated with non-family living. Despite the availability of data to support such an investigation, ours is the first study to examine changes over time in non-family living arrangements as well as the determinants of living with non-relatives. Our analysis was largely exploratory, and our results have both provided important empirical evidence around an increasingly common living arrangement in the United States and generated new questions for future research.

We found that non-family living is becoming more common in the United States. The share of individuals living with non-relatives only has increased over time at every age studied. Relative age trends have remained consistent, with the largest share of men and women in non-family living during the ages most associated with university attendance and then decreasing with age. Over time, the post-university decline in the share of young adults living in with non-family has flattened for both men and women but especially for men. The largest increase over time has been for the 25 to 29 age group. Consistent with studies of other living arrangements among young adults, this finding suggests more recent birth cohorts are delaying the formation of independent households.

The increase in non-family living arrangements among young adults is explained by changes in the composition and characteristics of the population over time. University education is the strongest predictor of non-family living for men and women ages 20 to 25 and 25 to 29. Viewed in the context of rapid educational expansion across the years captured in our sample, this finding might suggest that the increase in non-family living simply reflects the prolonging of third-level education as more young adults pursue postgraduate study and remain in student living arrangements. Indeed, school attendance is associated with non-family living, with a strong effect among 20- to 24-year-olds.

A more complex story emerges when we consider the other characteristics examined. Employment status is a stronger predictor of non-family living than school attendance at all ages for men and women. Living in a metropolitan area, being White, being foreign-born, and having a personal income below the poverty threshold are highly associated with non-family living for all age groups. These results indicate that non-family living is associated with markers of both advantage and disadvantage. On the one hand, having a four-year college degree, being employed, being White, and living in an urban area are characteristics associated with relative socio-economic advantage in the US. At the same time, the strong effects of poverty status suggest that the living arrangement is one of necessity rather than choice. Differences across age groups in the relative effect of control variables explain some of these mixed results and point to parallel profiles of individuals living with non-family that diverge with age. Our results suggest that non-family living is driven by economic necessity in urban areas among those not living with spouses, partners or children. This population is made up of single, childless people and people living apart from their families, and results are in line with our understanding of social and economic stratification of family formation throughout the life course.

At younger ages (20 to 24), the characteristics of those most likely to be living with non-family are in-line with the characteristics of people we expect to be single and childless: college-educated Whites. The strong effect of being college educated relative to other control variables and the large differences across racial and ethnic groups are consistent with what we know about socio-economic trends in family formation and living arrangements. Highly educated people cohabit, marry, and have children later than the less educated. White and Asian individuals are more likely to have a 4-year college degree than other racial and ethnic groups. Non-white individuals are more likely to live in multigenerational or complex family households and tend to marry/cohabit (Hispanic) and/or have children earlier (Black). These results point to one profile of individuals living in non-family arrangements: college-educated young white adults residing in urban areas earning relatively low wages as they begin careers. The poverty status variable used in our analysis does not capture intergenerational transfers of wealth or financial support received from other family members, which may be a significant source of finances among younger adults. The young adults living with non-family may have more financial resources available to them than what we are capturing in our models.

The foreign-born represent a parallel profile of those living with non-family. At all ages, the foreign-born are more likely than the native-born to live with non-family. At younger ages, the foreign-born may more closely resemble the overall population when it comes to the characteristics included in our model and their relationship with family formation, explaining the weaker effect of nativity relative to other characteristics like education level, school attendance, and income level for those aged 20 to 24. Among men aged 35 to 39, nativity is the strongest determinant of non-family living after marital status. At the same time, the relative effect of education, poverty status, and race change with age. This may reflect less pronounced social and economic stratification of family formation among older individuals. Most individuals are partnered and have children by the mid-30s, and gaps by educational level and race diminish greatly. These results suggest that at older ages, the foreign-born may be the least likely group to live with a partner, child, or other family members due to later marriage among foreign-born (Mayol-García et al. 2021) and the higher likelihood that the foreign-born live apart from spouses and children for periods of time. This interpretation is further supported by the fact that among those age 30 to 34 and 35 to 39, Hispanic men and Asian women are nearly as or more likely than white men and women to live with non-family only. The limitation of our analysis to capture all intrahousehold family relationships may be a contributing factor here as well. Previous studies (Van Hook et al. 2007) suggest that the foreign-born are more likely to live in complex households that include more distant family members—relationships we do not identify in all cases.

Educational expansion, delayed marriage and cohabitation, increased immigration and urbanization have increased the share of the population with the characteristics associated with non-family living. Trends among younger age groups are linked more closely to socio-economic patterns around family formation. Among older age groups, the demographic and labor force characteristics of the foreign born and constraints of their kin availability may be driving trends.

Our findings highlight some of the complexity of non-family living arrangements and generate a multitude of questions on the diversity of living arrangements among young adults in the United States and the implications of these trends. As this paper explores non-family living in isolation of other living arrangements, future investigations could consider non-family living alongside cohabitation with spouse or partner, parent-adult child coresidence, living alone, and residence in collective housing to provide a better understanding of how young adults across these living arrangements differ from one another. Individuals living in households where no one is related—"roommate" households—represent the majority of our sample, and results of our exploratory analysis are consistent whether or not individuals "hosted" by a family are also included. Future research could examine differences between these two groups of non-family dwellers. Studies could further explore aspects of these different types of non-family living arrangements: in "roommate" only households, do all household members have similar demographic and socio-economic characteristics? When an individual lives with a family that he or she is not related to, what are the relative situations of the "host" family and the non-family "guest" (like the work done by Harvey and Dunifon 2023 on households where parents and their children live with extended family). Another approach could be to compare results across a typology of complex households that contain at least one non-family member. Finally, foreign-born refers to an increasingly diverse population in the United States. Further study focused on this population is required to better determine what drives non-family living across migrant characteristics.

APPENDIX 3

Table A3.1 Sample characteristics and share in non-family households by age group, covariate, and census or survey year, males

	% of total sample				% in non-family households			
	20-24	25-29	30-34	35-39	20-24	25-29	30-34	35-39
University complete								
1990	10	22.3	24.6	28.5	23.6	15.1	6.1	3.5
2000	9.6	25.6	27.7	26.3	27	16.4	7.1	3.7
2010	11.2	27.1	29.5	30.2	25.7	18.4	7.6	3.6
2019	14.6	32.5	35.5	35.3	23.5	19.9	9.4	4.1
Attending school								
1990	30.9	13.7	8.8	6.7	7.9	9.1	5.4	3.9
2000	30.8	13.6	8	5.6	9.8	11.4	6.8	4.7
2010	37.3	15.4	8.6	5.6	10	13.1	7.9	5.5
2019	35	12.7	7.1	4.6	9.7	13.6	8.8	5.6
Employed								
1990	75.7	87	89.2	89.7	10.8	9.9	5.4	3.7
2000	73.5	83.2	85.3	85.5	12.4	12.1	6.7	4.4
2010	64	78.6	82.5	84	13.7	14.3	7.8	5.1
2019	72.5	84.7	87.9	88.4	13.3	14.8	8.9	5.4
Foreign-born								
1990	12.5	12.9	12.3	11.5	19.4	14.9	9.2	6.6
2000	17.9	20.7	20.2	17.8	20.7	16.6	10.4	7.4
2010	15.4	19.8	23.3	24.3	20.7	19.5	12	8.5
2019	11.8	15.2	20.2	23.2	21.3	19.1	11	7.8
Non-Hispanic white								
1990	71.4	74	76.2	77.9	15.8	9.8	5.1	3.5
2000	61.3	63	66.1	69.9	20.1	11.6	6	3.9
2010	57.3	59.6	59.0	60.4	21.2	13.4	6.8	4.3
2019	53.9	55.2	56.4	57.2	19.8	14.8	8.5	4.8
Single								
1990	77.3	44.2	24.3	14.7	18.4	19	15.4	13.6
2000	80.3	49.7	29.6	20.1	21.4	20.7	16.6	13.5
2010	89.5	63.8	38.5	25.3	20	19.4	15.4	13.2
2019	91.7	71.8	47.5	30.9	18.1	18.6	15.9	12.5

	% of total sample				% in non-family households			
	20-24	25-29	30-34	35-39	20-24	25-29	30-34	35-39
Living in metropolitan area								
1990	79.8	79.6	76.9	74.4	17.4	11.6	6.6	4.8
2000	78	80	80	78.5	18.9	13.4	7.8	5.3
2010	79.2	81.0	80.3	79.4	19	14.8	8.7	6.1
2019	81.5	83.6	84.0	82.4	17.5	15.3	9.6	6
Income below poverty threshold								
1990	17	10	8.7	7.8	38.8	17.9	11.4	9.6
2000	17.6	9.9	8.5	7.9	42.5	22.4	14.5	11.6
2010	22.4	13.5	11.9	10.8	43.7	23.2	15.4	12.4
2019	16.9	8.9	8.4	7.9	48.7	26.8	18.1	13.6

Source Authors's elaboration based on data from IPUMS USA

Table A3.2 Sample characteristics and share in non-family households by age group, covariate, and census or survey year, females

	% of total sample				% in non-family households			
	20-24	25-29	30-34	35-39	20-24	25-29	30-34	35-39
University complete								
1990	12.2	22.5	22.8	24.9	21.6	11.1	4.9	3
2000	13.7	29.9	29.2	26.3	22	10.9	4.4	2.6
2010	16.3	35	35.7	34.5	22.7	12.5	4.4	2.2
2019	21.1	39.9	42.9	42.4	21.9	13.9	5.4	2.8
Attending school								
1990	30.7	13.6	10.3	9.2	8.4	5.7	3.3	2.3
2000	34.6	15.1	9.3	7.1	9.4	6.3	3.4	2.7
2010	44.2	19.5	11.7	8.4	11.0	7.9	4	2.8
2019	42.2	16.1	9	6.5	12.6	9.5	4.9	3.6
Employed								
1990	66.5	69.9	69.4	72.2	13.2	7.5	4.1	2.7
2000	65.8	69.7	68.5	69.9	15.1	7.8	4.1	2.8
2010	62.3	69.7	69.4	69.8	16.7	9.5	4.4	2.8
2019	70.8	77.3	75.8	75.3	16.8	11.0	5.3	3.4
Foreign-born								
1990	10.5	11.4	11.3	11.2	12.2	7.9	5	3.7
2000	15.1	18.5	18.4	16.6	12.4	8.2	4.5	3.4
2010	13.3	18.7	22.7	24.1	14.2	8.9	5.1	3.5
2019	11.6	15.3	20.2	23.1	16.7	11.3	4.9	3.4
Non-Hispanic white								
1990	71.2	72.8	74.3	75.6	13.7	6.5	3.3	2.3
2000	61.6	62.4	65.1	68.3	16.7	7.4	3.5	2.6
2010	57.5	58.1	57.2	58.2	19.2	9.2	4.1	2.6
2019	53.1	54	55.3	55.6	20	11.1	5.2	3.3
Single								
1990	62.5	31.2	17.5	11.2	18	16	12.4	10.6
2000	68.8	38.2	21.7	14.9	19.2	15.1	11.3	9.8
2010	81.7	51.8	30.4	20.3	18.5	14.3	10	7.6
2019	86.4	61.5	37.8	25.5	18.4	15.1	10.5	8

	% of total sample				% in non-family households			
	20-24	25-29	30-34	35-39	20-24	25-29	30-34	35-39
Living in metropolitan area								
1990	80.6	79.3	76.9	75.3	14	7.3	4	2.8
2000	78.2	80.1	79.8	78.4	14.7	7.8	4	3
2010	79.8	81.1	80.4	79.8	16.2	9.3	4.6	2.9
2019	82.3	83.6	83.6	82.6	16.8	11	5.3	3.5
Income below poverty threshold								
1990	21	14.2	12.7	10.4	29.6	9.7	6.2	5.7
2000	24	14.6	12.6	11.3	30.6	11.2	7.15	6.7
2010	29.1	19.7	17.4	15.4	33.9	12.1	7.4	6.4
2019	21.9	14.6	13.5	12.9	41.2	16.8	9.5	8

Source Authors elaboration based on data from IPUMS USA

Table A3.3 Results of logistic regression of likelihood to live in a non-family household for males and females aged 20-24

	Odds Ratio (SE)			
	Males		Females	
Variable	*Model 1*	*Model 2*	*Model 1*	*Model 2*
Year				
1990 (reference)	1	1	1	1
2000	1.268 (.002)	1.325 (.002)	1.183 (.002)	1.014 (.002)
2010	1.277 (.002)	1.072 (.002)	1.350 (.002)	0.890 (.001)
2019	1.160 (.001)	1.060 (.002)	1.418 (.002)	1.030 (.002)
Educational attainment				
Primary or less (reference)		1		1
Secondary		1.416 (.002)		1.792 (.004)
University		2.742 (.006)		3.619 (.010)
School attendance		1.429 (.002)		1.585 (.002)
Employment status				
Employed (reference)		1		1
Unemployed		0.379 (.001)		0.456 (.001)
Inactive		0.553 (.001)		0.566 (.001)
Foreign-born		1.479 (.002)		1.162 (.002)
Race and ethnicity				
White (reference)		1		1
Non-Hispanic Black		0.344 (.001)		0.266 (0)
Non-Hispanic Asian		0.660 (.002)		0.864 (.002)
Non-Hispanic other		0.577 (.002)		0.602 (.002)
Hispanic any race		0.516 (.001)		0.427(.001)
Marital status				
Single (reference)		1		1
Married		0.101 (0)		0.085 (0)
Divorced, separated, widowed		1.044 (.004)		0.473 (.001)
Residence in metropolitan area		1.548 (.002)		1.665 (.002)
Poverty status				
Below threshold (reference)		1		1
Income 100-200% of threshold		0.327 (0)		0.341 (0)
Income > 200% of threshold		0.075 (0)		0.075 (0)

p < 0.001 for all variables

Source Authors' elaboration based on data from IPUMS USA

Table A4.4 Results of logistic regression of likelihood to live in a non-family household for males and females aged 30-34

	Odds Ratio (SE)			
	Males		Females	
Variable	*Model 1*	*Model 2*	*Model 1*	*Model 2*
Year				
1990 (reference)	1	1	1	1
2000	1.262 (.002)	1.124 (.002)	1.057 (.002)	0.926 (.002)
2010	1.449 (.003)	1.024 (.002)	1.219 (.003)	0.815 (.002)
2019	1.656 (.003)	1.078 (.002)	1.479(.003)	0.884 (.002)
Educational attainment				
Primary or less (reference)		1		1
Secondary		0.983 (.002)		1.033 (.003)
University		1.289 (.003)		1.508 (.005)
School attendance		1.157 (.003)		1.151 (.003)
Employment status				
Employed (reference)		1		1
Unemployed		0.626 (.002)		0.876 (.003)
Inactive		0.611 (.001)		0.730 (.002)
Foreign-born		1.757 (.003)		1.579 (.004)
Race and ethnicity				
White (reference)		1		1
Non-Hispanic Black		0.659 (.001)		0.389 (.001)
Non-Hispanic Asian		0.945 (.002)		1.071 (.004)
Non-Hispanic other		0.933 (.003)		0.838 (.004)
Hispanic any race		0.970 (.002)		0.707 (.002)
Marital status				
Single (reference)		1		1
Married		0.059 (0)		0.036 (0)
Divorced, separated, widowed		0.852 (.001)		0.565 (.001)
Residence in metropolitan area		1.680 (.003)		1.529 (.004)
Poverty status				
Below threshold (reference)		1		1
Income 100-200% of threshold		0.603 (.001)		0.700 (.002)
Income > 200% of threshold		0.349 (.001)		0.506 (.001)

$p < 0.001$ for all variable

Source Authors' elaboration based on data from IPUMS USA

Table A5.5 Results of logistic regression of likelihood to live in a non-family household for males and females aged 35-39

	Odds Ratio (SE)			
	Males		Females	
Variable	*Model 1*	*Model 2*	*Model 1*	*Model 2*
Year				
1990 (reference)	1	1	1	1
2000	1.198 (.003)	1.021 (.002)	1.123 (.003)	1.001 (.003)
2010	1.408 (.003)	0.957 (.002)	1.134 (.003)	0.811 (.002)
2019	1.440 (.003)	0.956 (.002)	1.400 (.004)	0.948 (.003)
Educational attainment				
Primary or less (reference)		1		1
Secondary		0.888 (.002)		0.956 (.003)
University		0.878 (.002)		1.100 (.004)
School attendance		1.248 (.004)		1.029 (.003)
Employment status				
Employed (reference)		1		1
Unemployed		0.742 (.002)		0.859 (.004)
Inactive		0.667 (.002)		0.835 (.002)
Foreign-born		1.885 (.004)		1.518 (.005)
Race and ethnicity				
White (reference)		1		1
Non-Hispanic Black		0.802 (.002)		0.426 (.001)
Non-Hispanic Asian		0.905 (.004)		1.176 (.005)
Non-Hispanic other		0.880 (.004)		0.790 (.005)
Hispanic any race		1.103 (.003)		0.707 (.002)
Marital status				
Single (reference)		1		1
Married		0.058 (0)		0.037 (0)
Divorced, separated, widowed		0.874 (.002)		0.613 (.001)
Residence in metropolitan area		1.655 (.004)		1.275 (.003)
Poverty status				
Below threshold (reference)		1		1
Income 100-200% of threshold		0.586 (.001)		0.612 (.002)
Income > 200% of threshold		0.387 (.001)		0.488 (.001)

p < 0.001 for all variables

Source Authors' elaboration based on data from IPUMS USA

A TALE OF TWO SOURCES: CONCLUSIONS

This thesis examined how harmonized population microdata shape our understanding of recent demographic change by assessing coherence across data sources and exploring applications of integrated census and survey microdata. Across three essays, I contributed new knowledge in several ways. First, I documented trends in female educational attainment using available data from IPUMS International, DHS, and MICS and assessed coherence across sources. In the next chapter, I identified discrepancies in household size and composition between IPUMS International census samples and DHS surveys and, building on themes explored in the first essay, shed light on their origins. These investigations of coherence across data sources are among the first of their kind, providing new and practical insights to the research community. Finally, I exploited integrated U.S. census and ACS data to describe trends in non-family living arrangements among young adults, an area of demographic inquiry with limited prior research. United by a reliance on IPUMS data, the chapters combine methodological scrutiny with substantive inquiry to strengthen evidence on the reliability of comparative research and illustrate the analytic potential of pooled microdata. Implications for data collection, analysis, and future research are discussed in the following paragraphs.

SUMMARY OF FINDINGS

CHAPTER 1: TRENDS IN FEMALE EDUCATION IN LOW- AND MIDDLE-INCOME COUNTRIES: COHERENCE ACROSS DATA SOURCES

The essay presented in chapter 1 evaluated whether three major sources of international microdata—IPUMS International, DHS, and MICS—provide a coherent account of female educational expansion in LMICs. Drawing on data from 535 censuses and surveys from 75 countries, the essay documents the broad rise in secondary completion among women aged 25 to 29 across regions while systematically assessing where and when census and survey observations diverge. The results indicate generally high consistency across sources but also identify notable misalignments. Thirty percent of observed point estimates diverge from predicted trends by two or more percentage points. Point estimates derived from censuses align with predicted trends better than estimates from DHS and MICS, and coherence across data sources is strongest in Africa and lowest in Latin America.

The findings have practical applications and implications. The adjusted estimates of educational attainment produced to gauge the alignment of data sources can be used in other studies, serving as benchmarks to assess alignment of future censuses and DHS and MICS surveys or to evaluate the quality of measures of education based on other data sources. The study also identified specific samples that warrant caution or further investigation of potential biases, including those related to the sampling frames used in surveys and the processing of education data in both censuses and surveys.

CHAPTER 2: DISCREPANCIES IN HOUSEHOLD SIZE AND COMPOSITION BETWEEN CENSUS AND DHS DATA ACROSS 22 LOW- AND MIDDLE-INCOME COUNTRIES

The essay presented in chapter 2 explored coherence in measures of household size and composition across DHS surveys and census data available from IPUMS International in 22 LMICs. The first phase of the analysis compared average household size in 124 censuses samples and DHS surveys and found that in three-quarters of country-year observations, DHS surveys report larger households than those captured in census data. On average, the absolute difference in household size between the two sources is 0.31 persons; in 20 percent of observations the absolute is 0.5 persons. Further investigation suggested inconsistencies in the average number of adult women and children across data sources drive discrepancies in household size. Individual-level analyses of microdata for Benin, Guinea, Senegal and Tanzania revealed differences in fertility and characteristics linked to fertility between data sources. A larger share of women living without children are reported in the census samples than in the DHS surveys in all four countries. In Benin, Senegal, and Tanzania, the share of single women is higher in the the census samples than in the DHS surveys.

These findings have important implications for data analysis and collection. As a primary source of data for fertility estimates, the potential underrepresentation of childless women in DHS samples should be studied in future research, for example by comparing indirect measures of fertility across DHS surveys and censuses. The bias introduced by deviations from sampling frames should also be explored. Adding to the evidence on data coherence generated in chapter 1, the results equip researchers with practical information to help them responsibly design studies and accurately interpret results based on pooled census and DHS data.

CHAPTER 3: NON-FAMILY LIVING ARRANGEMENTS AMONG YOUNG ADULTS IN THE UNITED STATES

The essay presented in chapter 3 documented trends over time in non-family living arrangements among young adults in the United States and examined the sociodemographic profile of those living with non-relatives. An analysis of pooled U.S. census and ACS microdata for the period 1990 to 2019 indicated that the share of individuals living with non-family has increased over time across all age groups studied. Models estimating the likelihood of living with non-family based on individual-level characteristics suggest that the arrangement is associated with markers of both advantage and disadvantage and that there are differences across age groups. Results provide evidence to support the idea that non-family living is driven by economic necessity in urban areas among single, childless people and people living apart from their families. Patterns among younger age groups are connected to socioeconomic differences in family formation: white, college-educated young adults tend to postpone cohabitation, marriage, and childbearing and are also more likely to reside in non-family households. Among older age groups, the demographic and employment characteristics of the foreign-born, along with limitations in kin availability, may be shaping observed trends.

The findings add to the literature on living arrangements of young adults, expanding the evidence beyond cohabitation among unmarried couples and parent-adult child coresidence. The paper represents an application of pooled census and survey data from IPUMS to study a rare and complex living arrangement and identifies several areas for future research, including work that examines how the characteristics of young adults differ by living arrangement, that explores the demographic and socioeconomic situation of these non-family households, and that further investigates non-family living among migrants.

DISCUSSION

This thesis provides empirical evidence to support a notion that researchers have long understood–the data we use determine the story we tell. In the current era of "big population microdata" (Ruggles 2014), the range of possible stories has expanded dramatically, making it all the more important to understand how data availability, quality, and comparability influence empirical findings. As demonstrated throughout this thesis, even nationally representative data from traditional sources are not immune to bias. These biases are statistical–like those introduced by sampling error–but they are also introduced by humans–the humans that design, collect, process, and interpret the data.

Demographers continue to embrace new sources of data, and these data will bring new biases. Studies increasingly integrate geospatial, biosocial, administrative, and “digital trace” data (Kashyap 2021). Recently available genetic data may open up new lines of inquiry exploring the interplay of genetics, environmental exposure, and demographic outcomes (Breen and Feehan 2025). Novel approaches to sampling are also generating new datasets on small and hard-to-reach population groups (Breen and Feehan 2025). Research with new sources will rely on traditional data sources like population censuses and sample surveys to shed light on biases--to validate samples, calibrate models, generate weights–and to provide data on key demographic outcomes (Breen and Feehan 2025).

In my current job with the United Nations Economic Commission for Europe, however, I see changing attitudes towards traditional methodologies in official statistics. The census has entered a new phase shaped by digital technology, administrative data, and concerns about cost, privacy, and data quality. Already growing demand for more timely data was exacerbated by the real-time information needs of the Covid-19 pandemic. National Statistical Offices are moving away from complete, field-based census enumerations to combined methods using administrative data or fully register-based systems that can produce census-like data more frequently at a lower cost and with reduced burden to respondents. At the same time, the political appetite for specialized sample surveys is waning amidst decreasing response rates, high costs, and stretched budgets. Growing cyber threats have heightened public sensitivity towards the dissemination of personal data and led to restrictions around the dissemination of individual-level data–even anonymized–in some jurisdictions.

The challenges are different in LMICs, with resource and capacity constraints affecting traditional data collection efforts. Recent cuts by the United States in foreign aid and contributions to international organizations has created uncertainty around the future of international survey programs like the Demographic and Health Surveys and the UNICEF MICS surveys. A potential outcome of this shifting data landscape is reduced availability of traditional data from censuses and surveys. While new sources may be used in their place, the need for representative population-level data for reliable demographic research will remain. If these data become less common, researchers will continue to find innovative ways to work with existing data, intregrating data from multiple sources to address gaps and validate new sources. In this context of new data and methods, there may be an inclination among researchers to assume that compared to emerging sources, traditional data sources are inherently reliable. But as traditional sources are put to use in new ways and looked to as "ground truth" for new sources, it is critical to continuously assess their reliability to ensure that these "old" data are fit for new purposes.

The findings of this thesis suggest that key demographic measures based on DHS and MICS survey data are sufficiently coherent with those derived from census data as not to preclude pooled analyses. At the same time, the results provide evidence to justify the need for careful interpretation of results based on pooled data related to a number of key demographic domains including educational attainment, household size and composition, and various fertility-related determinants and outcomes. While the outcomes of this thesis have strengthened my trust in survey data as a complement to census microdata, the results also reinforce the unique merit of the census as the only information source on the entire population in many countries. As countries continue to innovate census operations in the future, the core purpose of the census should remain consistent: to provide a comprehensive record of a country's population, its characteristics and living conditions. The demographic data revolution will continue to evolve, but census data will endure as the foundation of demographic study and evidence-based policymaking.

LIMITATIONS

This thesis has several limitations that extend beyond those discussed in the individual essays. First, although the analyses draw on three of the most widely used sources of population microdata—IPUMS, DHS, and MICS—these are not the only infrastructures that underpin comparative demographic research or that are pooled by researchers. Important data sources such as the European Union Survey on Income and Living Conditions (EU-SILC), the European Union Labour Force Survey (EU-LFS), the Harmonised European Time Use Surveys (HETUS), the Generations and Gender Survey, national population registers, and country-specific household surveys fall outside the scope of this dissertation. As a result, the findings speak only to coherence across select data sources and cannot be generalized to all forms of population data used in comparative research.

A second limitation concerns the inherent heterogeneity of data quality across countries, censuses, and surveys. Each data collection effort is shaped by its own institutional, logistical, and political context, including the recency and accuracy of sampling frames, the geographic, cultural, and linguistic complexity of fieldwork, and the specific design and implementation choices made by National Statistical Offices and survey organizations. These idiosyncratic sources of bias are not comprehensively documented and could not be fully evaluated within the scope of this thesis. A systematic assessment of country-specific data quality would require a level of investigation far beyond what was feasible here.

Finally, the thesis is constrained by the timing of data availability. As with most quantitative research on contemporary demographic change, the evidence presented here reflects the data landscape at the time the analyses were conducted. New census microdata continue to be published, and IPUMS is rapidly expanding its harmonized collections. Funding for DHS and MICS appears secure in the near term, suggesting that new data will continue to be collected and released. While the inclusion of recently released samples would likely not alter the core conclusions of the thesis, the topics examined would undoubtedly benefit from broader geographic coverage and more recent observations. The planned expansion of IPUMS DHS to cover other world regions (King 2025) and the launch of IPUMS MICS in particular opens up many new areas of inquiry. Continued data dissemination will allow future research to refine, extend, and update the findings presented here.

REFERENCES

Abuya, B. A., Ciera, J., & Kimani-Murage, E. (2012). Effect of mother's education on child's nutritional status in the slums of Nairobi. *BMC Pediatrics, 12*(1), 80.

Abuya, B. A., Onsomu, E. O., Kimani, J. K., & Moore, D. (2011). Influence of maternal education on child immunization and stunting in Kenya. *Maternal and Child Health Journal, 15*(8), 1389–1399.

Ackah, J. A., Sear, R., McLean, E., Awusabo-Asare, K., Hassan, A., Achana, F. S., & Walters, S. (2025). Household structure in Ghana: Exploring dynamics over three decades (No. 7dhqr_v1). Center for Open Science.

Ahmed, S., Li, Q., Scrafford, C., & Pullum, T. W. (2014). *An assessment of DHS maternal mortality data and estimates* (DHS Methodological Reports No. 13). ICF International. http://dhsprogram.com/pubs/pdf/MR13/MR13.pdf

Ainsworth, M., Beegle, K., & Nyamete, A. (1996). The impact of women's schooling on fertility and contraceptive use: A study of fourteen sub-Saharan African countries. *The World Bank Economic Review, 10*(1), 85–122.

Alkema, B. (2020). The demography of Sub-Saharan Africa in the 21st century. *Population, 75*(2), 165–286.

Allen, C. K., Fleuret, J., & Ahmed, J. (2020). *Data quality in Demographic and Health Surveys that used long and short questionnaires* (DHS Methodological Reports No. 30). ICF. https://www.dhsprogram.com/pubs/pdf/MR30/MR30.pdf

Amorim, M., Dunifon, R., & Pilkauskas, N. (2017). The magnitude and timing of grandparental coresidence during childhood in the United States. *Demographic Research, 37*, 1695–1706.

Barakat, B. F., & Durham, R. E. (2014). Future education trends. In *World Population and Human Capital in the 21st Century* (pp. 397–433). Oxford University Press.

Barro, R. J., & Lee, J. W. (1994, June). Sources of economic growth. In *Carnegie-Rochester Conference Series on Public Policy* (Vol. 40, pp. 1–46). North-Holland.

Barro, R. J., & Lee, J. W. (2013). A new data set of educational attainment in the world, 1950–2010. *Journal of Development Economics, 104*, 184–198.

Batyra, E., & Pesando, L. M. (2021). Trends in child marriage and new evidence on the selective impact of changes in age-at-marriage laws on early marriage. *SSM–Population Health, 14*, 100811.

Bauer, R., Potančoková, M., Goujon, A., & KC, S. (2012). *Populations for 171 countries by age, sex, and level of education around 2010: Harmonized estimates of the baseline data for the Wittgenstein Centre projections.* IIASA Interim Report. IIASA, Laxenburg, Austria: IR-12-016

Becker, G. S. (1981). *A treatise on the family* (Enlarged ed.). Harvard University Press.

Becker, G. S. (1994). *Human capital: A theoretical and empirical analysis, with special reference to education.* University of Chicago Press.

Bicego, G. T., & Boerma, J. T. (1993). Maternal education and child survival: A comparative study of survey data from 17 countries. *Social Science & Medicine, 36*(9), 1207–1227.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.). (2013). *Measurement errors in surveys.* John Wiley & Sons.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality.* Wiley.

Bongaarts, J. (2001). Household size and composition in the developing world in the 1990s. *Population Studies, 55*(3), 263–279.

Boyle, E. H., King, M., & Sobek, M. (2020). *IPUMS-Demographic and Health Surveys: Version 8* [dataset]. IPUMS and ICF. https://doi.org/10.18128/D080.V8

Boyle, E. H., King, M., & Sobek, M. (2024). *IPUMS-Demographic and Health Surveys: Version 11* [Dataset]. IPUMS & ICF. https://doi.org/10.18128/D080.V11

Breen, C. F., & Feehan, D. M. (2025). New data sources for demographic research. *Population and Development Review*, *51*(1), 539-573.

Caldwell, J. C. (1981). Maternal education as a factor in child mortality. *World Health Forum, 2*(1), 75–78.

Caldwell, J., & McDonald, P. (1982). Influence of maternal education on infant and child mortality: Levels and causes. *Health Policy and Education, 2*(3–4), 251–267.

Castro, T. (1995). Women's education and fertility: Results from 26 Demographic and Health Surveys. *Studies in Family Planning*, 187–202.

Cherlin, A. J. (2010). Demographic trends in the United States: A review of research in the 2000s. *Journal of Marriage and Family, 72*(3), 403–419.

Christian, P. B. (1989). Nonfamily households and housing among young adults. In *Ethnicity and the new family economy* (pp. 57–73). Routledge.

Cochrane, S. H., Leslie, J., & O'Hara, D. J. (1982). Parental education and child health: Intracountry evidence. *Health Policy and Education, 2*(3–4), 213–250.

Cohen, D., & Soto, M. (2007). Growth and human capital: Good data, good results. *Journal of Economic Growth, 12*(1), 51–76.

Cross, C. (2018). Extended family households among children in the United States: Differences by race/ethnicity and socio-economic status. *Population Studies, 72*, 235–251.

De la Fuente, A., & Domenech, R. (2013). Cross-country data on the quantity of schooling: A selective survey and some quality measures. *BBVA Working Papers, 13/27*. Madrid.

Desai, S., & Alva, S. (1998). Maternal education and child health: Is there a strong causal relationship? *Demography, 35*(1), 71–81.

Di Giulio, P., Impicciatore, R., & Sironi, M. (2019). The changing pattern of cohabitation: A sequence analysis approach. *Demographic Research*, *40*, 1211–1248.

Dorius, S. F. (2013). The rise and fall of worldwide education inequality from 1870 to 2010: Measurement and trends. *Sociology of Education, 86*(2), 158–173.

Dorius, S. F., & Firebaugh, G. (2010). Trends in global gender inequality. *Social Forces, 88*(5), 1941–1968.

Elksabi, M., & Khan, A. (2022). *Modeling interviewer effects in DHS surveys* (DHS Methodological Reports No. 32). ICF. https://www.dhsprogram.com/pubs/pdf/MR32/MR32.pdf

Esteve, A., Pohl, M., Becca, F., Fang, H., Galeano, J., García-Román, J., ... & Turu, A. (2024). A global perspective on household size and composition, 1970–2020. *Genus, 80*(1), 2.

Esteve, A., Reher, D. S., Treviño, R., Zueras, P., & Turu, A. (2020). Living alone over the life course: Cross-national variations on an emerging issue. *Population and Development Review, 46*(1), 169–189.

Esteve, A., & Reher, D. S. (2021). Rising global levels of intergenerational coresidence among young adults. *Population and Development Review, 47*(3), 691–717.

Esteve, A., & Reher, D. S. (2024). Trends in living arrangements around the world. *Population and Development Review, 50*(1), 211–232.

Esteve, A., & Ruggles, M. (2003). Challenges and methods of international census harmonization. *Historical Methods*, *36*, 66–70.

Esteve, A., Schwartz, C. R., Van Bavel, J., Permanyer, I., Klesment, M., & Garcia, J. (2016). The end of hypergamy: Global trends and implications. *Population and Development Review, 42*(4), 615–625. https://doi.org/10.1111/padr.12012

Fukuda-Parr, S., Greenstein, J., & Stewart, D. (2013). How should MDG success and failure be judged: Faster progress or achieving the targets? *World Development, 41*, 19–30.

Furstenberg, F. F. (2008). The intersections of social class and the transition to adulthood. In J. T. Mortimer (Ed.), *Social class and transitions to adulthood* (pp. 1–10). Wiley.

Furstenberg, F. F. (2010). On a new schedule: Transitions to adulthood and family change. *The Future of Children, 20*(1), 67–87.

Furstenberg, F. F. (2015). Will marriage disappear? *Proceedings of the American Philosophical Society, 159*(3), 241–246.

Furstenberg, F. F. (2019). Family change in global perspective: How and why family systems change. *Family Relations, 68*(3), 326–341.

Galeano, J., Esteve, A., Turu, A., et al. (2024). CORESIDENCE: National and subnational data on household size and composition around the world, 1964–2021. *Scientific Data, 11*, 145. https://doi.org/10.1038/s41597-024-02964-3

Goujon, A., KC, S., Speringer, M., Barakat, B., Potančoková, M., Eder, J., Striessnig, E., Bauer, R., & Lutz, W. (2016). A harmonized dataset on global educational attainment between 1970 and 2060: An analytical window into recent trends and future prospects in human capital development. *Journal of Demographic Economics, 82*(3), 315–363.

Grant, M. J., & Behrman, J. R. (2010). Gender gaps in educational attainment in less developed countries. *Population and Development Review, 36*(1), 71–89.

Grépin, K. A., & Bharadwaj, P. (2015). Maternal education and child mortality in Zimbabwe. *Journal of Health Economics, 44*, 97–117.

Hannum, E., & Buchmann, C. (2005). Global educational expansion and socio-economic development: An assessment of findings from the social sciences. *World Development, 33*(3), 333–354.

Harvey, H., & Dunifon, R. (2023). Why mothers double up: The role of demographic, economic, and family characteristics. *Journal of Marriage and Family, 85*(3), 845–868.

Heath, S., & Kenyon, L. (2001). Single young professionals and shared household living. *Journal of Youth Studies, 4*(1), 83–100.

Hemez, P., & Manning, W. D. (2017). Twenty-five years of change in cohabitation experience in the U.S., 1987–2013. *Family Profiles, FP-17-02*. Bowling Green, OH: National Center for Family & Marriage Research.

Hertrich, V., Lardoux, S., & George, K. (2014). Estimating age at first union in Africa: Are census and survey data comparable? *Population, 69*(3), 357–389.

Herron, J. (2019, February 9). Help with the mortgage: More married couples bring in roommates to ease cost, study shows. *USA Today*. https://www.usatoday.com/story/money/2019/02/08/personal-finance-why-more-married-couples-have-roommates/2812388002/

Houle, J. N., & Warner, C. (2017). Into the red and back to the nest? Student debt, college completion, and returning to the parental home among young adults. *Sociology of Education, 90*(1), 89–108.

ICF. (2004-2017). Demographic and Health Surveys (various) [Datasets]. Funded by USAID. Rockville, Maryland: ICF [Distributor].

ICF. (2019). Survey types. *The DHS Program*. http://www.dhsprogram.com

ICF International. (2012). *Demographic and Health Surveys methodology: Sampling and household listing manual.* MEASURE DHS, Calverton, MD: ICF International.

Ilie, S., & Rose, P. (2016). Is equal access to higher education in South Asia and sub-Saharan Africa achievable by 2030? *Higher Education, 72*(4), 435–455.

IPUMS. (2026). *IPUMS by the numbers*. https://www.ipums.org/by-the-numbers

IPUMS International. (2026a). *Age-sex reporting assessment.* https://international.ipums.org/international/age_reporting.shtml

IPUMS International. (2026b). *IPUMS International*. https://international.ipums.org/international/

Jacob, A. (2017). Mind the gap: Analyzing the impact of data gap in Millennium Development Goals' indicators on the progress toward MDGs. *World Development, 93*, 260–278.

Jeong, J., Bhatia, A., & Fink, G. (2018). Associations between birth registration and early child growth and development: Evidence from 31 low- and middle-income countries. *BMC Public Health, 18*(1), 673.

Jeffers, K., & Esteve, A. (2022). Trends in female education in low- and middle-income countries: Coherence across data sources. *Comparative Population Studies, 47*.

Johnson, P., Spoorenberg, T., Hertog, S., & Gerland, P. (2022). *Method protocol for the evaluation of census population data by age and sex* (No. 5). UN DESA.

Jordá, V., & Alonso, J. M. (2017). New estimates on educational attainment using a continuous approach (1970–2010). *World Development, 90*, 281–293.

Kamo, Y. (2000). Racial and ethnic differences in extended family households. *Sociological Perspectives, 43*(2), 211–229.

Kang, Y., Aguayo, V. M., Campbell, R. K., & West, K. P., Jr. (2018). Association between stunting and early childhood development among children aged 36–59 months in South Asia. *Maternal & Child Nutrition, 14*, e12684.

Kashyap, R. (2021). Has demography witnessed a data revolution? Promises and pitfalls of a changing data ecosystem. *Population Studies*, *75*(sup1), 47–75.

Kc, S., Barakat, B., Goujon, A., Skirbekk, V., Sanderson, W. C., & Lutz, W. (2010). Projection of populations by level of educational attainment, age, and sex for 120 countries for 2005–2050. *Demographic Research, 22*(15), 383–472.

Kc, S., & Lentzner, H. (2010). The effect of education on adult mortality and disability: A global perspective. *Vienna Yearbook of Population Research*, 201–235.

Kebede, E., Goujon, A., & Lutz, W. (2019). Stalls in Africa's fertility decline partly result from disruptions in female education. *Proceedings of the National Academy of Sciences, 116*(8), 2891–2896.

Keene, J. R., & Batson, C. D. (2010). Under one roof: A review of research on intergenerational coresidence and multigenerational households in the United States. *Sociology Compass, 4*(8), 642–657.

King, M. (2025). Demographic and Health Survey (DHS) data again available to new users. *Use it for good: News and information from IPUMS*. https://blog.popdata.org/dhs-data-again-available-to-new-users/

Kobrin, F. E. (1976). The fall in household size and the rise of the primary individual in the United States. *Demography, 13*(1), 127–138.

Kuperberg, A. (2014). Age at coresidence, premarital cohabitation, and marriage dissolution: 1985–2009. *Journal of Marriage and Family, 76*(2), 352–369.

Lei, L., & South, S. J. (2016). Racial and ethnic differences in leaving and returning to the parental home: The role of life course transitions, socioeconomic resources, and family connectivity. *Demographic Research, 34*, 109.

Lesthaeghe, R. (2014). The second demographic transition: A concise overview of its development. *Proceedings of the National Academy of Sciences, 111*(51), 18112–18115.

Lesthaeghe, R. J., & Neidert, L. (2006). The second demographic transition in the United States: Exception or textbook example? *Population and Development Review*, 669–698.

Liu, C., & Esteve, A. (2020). *Living arrangements across households in Europe* (No. WP-2020-002). Max Planck Institute for Demographic Research.

Lloyd, C. (2005). *Growing up global: The transition to adulthood in developing countries.* National Academies Press.

Lutz, W., & Skirbekk, V. (2014). How education drives demography and knowledge informs projection. In *World Population and Human Capital in the 21st Century* (pp. 14–38). Oxford University Press.

Lutz, W., Butz, W. P., & KC, S. (Eds.). (2014). *World population and human capital in the 21st century.* Oxford University Press.

Lutz, W., Goujon, A., Kc, S., Stonawski, M., & Stilianakis, N. (2018). *Demographic and human capital scenarios for the 21st century: 2018 assessment for 201 countries.* Publications Office of the European Union.

Magnuson, D. L., & Ruggles, S. (2022). Challenges of large-scale data processing in the 1990s: The IPUMS experience. *IEEE Annals of the History of Computing*, *44*(4), 71–83.

Magnuson, D. L., & Ruggles, S. (2025). "Equally Available to Anyone": The Invention of Public Use Microdata at the US Census Bureau, Version 2. IPUMS Working Paper No. 2024-07.

Makoka, D., & Masibo, P. K. (2015). Is there a threshold level of maternal education sufficient to reduce child undernutrition? Evidence from Malawi, Tanzania and Zimbabwe. *BMC Pediatrics, 15*(1), 96.

Manning, W. D. (2020). Thirty years of change in cohabitation experience in the U.S., 1987–2017. *Family Profiles, FP-20-27*. Bowling Green, OH: National Center for Family & Marriage Research.

Manning, W. D., Brown, S. L., & Payne, K. K. (2014). Two decades of stability and change in age at first union formation. *Journal of Marriage and Family, 76*(2), 247–260.

Manning, W. D., & Cohen, J. A. (2015). Teenage cohabitation, marriage, and childbearing. *Population Research and Policy Review, 34*(2), 161–177.

Masquelier, B., & Garbero, A. (2016). Educational differentials in adult mortality in low- and middle-income countries. *Quetelet Journal, 4*(1), 7–28.

Mather, M., Jacobsen, L., Jarosz, B., Kilduff, L., Lee, A., Pollard, K., Scommegna, P., & Vanorman, A. (2019). What the 2020 Census will tell us about a changing America. *Population Bulletin, 74*(1).

Matsudaira, J. D. (2016). Economic conditions and the living arrangements of young adults: 1960 to 2011. *Journal of Population Economics, 29*(1), 167–195.

Mayol-García, Y. H., Gurrentz, B., & Kreider, R. M. (2021). Number, timing, and duration of marriages and divorces: 2016. *Current Population Reports, P70–167*. U.S. Census Bureau.

Mazurik, K., Knudson, S., & Tanaka, Y. (2020). Stuck in the nest? A review of the literature on coresidence in Canada and the United States. *Marriage & Family Review, 56*(6), 491–512.

McCaa, R., & Ruggles, S. (2002). The census in global perspective and the coming microdata revolution. *Scandinavian Population Studies*, *13*, 7-30.

McLanahan, S. (2004). Diverging destinies: How children are faring under the second demographic transition. *Demography, 41*(4), 607–627.

Mellington, N., & Cameron, L. (1999). Female education and child mortality in Indonesia. *Bulletin of Indonesian Economic Studies, 35*(3), 115–144.

Miller, G. H. (2020). Future trends in marriage and family. In *International handbook on the demography of marriage and the family* (pp. 307–315). Springer.

Minnesota Population Center. (2020). *Integrated Public Use Microdata Series, International: Version 7.3* [dataset]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D020.V7.3

Mukherjee, B. N., & Mukhopadhyay, B. K. (1988). A study of digit preference and quality of age data in Turkish censuses. *Genus*, 201–227.

Moylan, M. (2019, July 10). Taking a roommate can make economic sense for homeowners, tenants. *Minnesota Public Radio News*. https://www.mprnews.org/story/2019/07/10/taking-a-roommate-can-make-economic-sense-for-homeowners-tenants

Nagi, M. H., Stockwell, E. G., & Snavley, L. M. (1973). Digit preference and avoidance in the age statistics of some recent African censuses: Some patterns and correlates. *International Statistical Review, 41*(2), 165–174.

OECD. (2018). *Education at a glance 2018: OECD indicators.* OECD Publishing. https://doi.org/10.1787/eag-2018-en

Osterman, M. J. K., Hamilton, B. E., Martin, J. A., Driscoll, A. K., & Valenzuela, C. P. (2022). *Births: Final data for 2020* (National Vital Statistics Reports, Vol. 70, No. 17). National Center for Health Statistics. https://www.cdc.gov/nchs/data/nvsr/nvsr70/nvsr70-17.pdf

Pace, G. T., Lee, S. J., & Grogan-Kaylor, A. (2019). Spanking and young children's socioemotional development in low- and middle-income countries. *Child Abuse & Neglect, 88*, 84–95.

Pappu, S. (2016, May 5). Age 31 and up, with roommates. You got a problem with that? *The New York Times*. https://www.nytimes.com/2016/05/06/fashion/mens-style/adult-men-roommates-new-york.html

Pohl, M., Esteve, A., & Galeano, J. (2025). African households: National and subnational trends from censuses and surveys. *Population Studies*, 1–21.

Pullum, T. W. (2019). *Strategies to assess the quality of DHS data* (DHS Methodological Reports No. 26). ICF. http://dhsprogram.com/pubs/pdf/MR26/MR26.pdf

Pullum, T. W., & Becker, S. (2014). *Evidence of omission and displacement in DHS birth histories* (DHS Methodological Reports No. 11). ICF International. http://dhsprogram.com/pubs/pdf/MR11/MR11.pdf

Pullum, T. W., & Staveteig, S. (2017). *An assessment of the quality and consistency of age and date reporting in DHS surveys, 2000–2015* (DHS Methodological Reports No. 19). ICF. http://dhsprogram.com/pubs/pdf/MR19/MR19.pdf

Pullum, T. W., Assaf, S., & Staveteig, S. (2017). *Comparisons of DHS estimates of fertility and mortality with other estimates* (DHS Methodological Reports No. 21). ICF. http://dhsprogram.com/pubs/pdf/MR21/MR21.pdf

Pullum, T. W., Juan, C., Khan, N., & Staveteig, S. (2018). *The effect of interviewer characteristics on data quality in DHS surveys* (DHS Methodological Reports No. 24). ICF. http://dhsprogram.com/pubs/pdf/MR24/MR24.pdf

Randall, S., & Coast, E. (2016). The quality of demographic data on older Africans. *Demographic Research, 34*, 143–174.

Rerimoi, A. J., Jasseh, M., Agbla, S. C., Reniers, G., Roca, A., & Timæus, I. M. (2019). Under-five mortality in The Gambia: Comparison of the results of the first demographic and health survey with those from existing inquiries. *PLoS ONE, 14*(7), e0219919.

Reyes, A., Schoeni, R. F., & Choi, H. (2020). Race/ethnic differences in spatial distance between adult children and their mothers. *Journal of Marriage and Family, 82*(2), 810–821.

Rosenzweig, M. R., & Wolpin, K. I. (1994). Are there increasing returns to the intergenerational production of human capital? Maternal schooling and child intellectual achievement. *The Journal of Human Resources, 29*(2), 670–693.

Rosenthal, M. J. (2007). *The age of independence: Interracial unions, same-sex unions, and the changing American family.* Harvard University Press.

Ruggles, Steven. 2014. Big microdata for population research, *Demography 51*(1): 287–297.

Ruggles, S., Flood, S., Foster, S., Goeken, R., Pacas, J., Schouweiler, M., & Sobek, M. (2021). *IPUMS USA: Version 11.0* [Dataset]. IPUMS. https://doi.org/10.18128/D010.V11.0

Ruggles, S., Cleveland, L., Lovaton, R., Sarkar, S., Sobek, M., Burk, D., Ehrlich, D., Heimann, Q., Lee, J., & Merrill, N. (2025). *Integrated Public Use Microdata Series, International: Version 7.6* [Dataset]. IPUMS. https://doi.org/10.18128/D020.V7.6

Ruggles, S., Hacker, J. D., & Sobek, M. (1995). General design of the integrated public use microdata series. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *28*(1), 33-39.

Rutstein, S. O. (2014). *Potential bias and selectivity in analyses of children born in the past five years using DHS data* (DHS Methodological Reports No. 14). ICF International. http://dhsprogram.com/pubs/pdf/MR14/MR14.pdf

Sandberg-Thoma, S. E., Snyder, A. R., & Jang, B. J. (2015). Exiting and returning to the parental home for boomerang kids. *Journal of Marriage and Family, 77*(3), 806–818.

Sassler, S., & Lichter, D. T. (2020). Cohabitation and marriage: Complexity and diversity in union-formation patterns. *Journal of Marriage and Family, 82*(1), 35–61.

Sassler, S., & Miller, A. (2017). *Cohabitation nation.* University of California Press.

Sassler, S., Michelmore, K., & Qian, Z. (2018). Transitions from sexual relationships into cohabitation and beyond. *Demography, 55*(2), 511–534.

Schady, N. (2011). Parents' education, mothers' vocabulary, and cognitive development in early childhood: Longitudinal evidence from Ecuador. *American Journal of Public Health, 101*(12), 2299–2307.

Schoumaker, B. (2014). *Quality and consistency of DHS fertility estimates, 1990 to 2012* (DHS Methodological Reports No. 12). ICF International. http://dhsprogram.com/pubs/pdf/MR12/MR12.pdf

Shipanga, I., & Shinyemba, T. (2023). Evaluation of age-sex data collected in the 1991, 2001 and 2011 Population and Housing Censuses of Namibia. *Asian Journal of Population Sciences*, 1–16.

Singh, K., Li, Q., Ahsan, K. Z., Curtis, S., & Weiss, W. (2022). A comparison of approaches to measuring maternal mortality in Bangladesh, Mozambique, and Bolivia. *Population Health Metrics, 20*(1), 5.

Singh, M., Kashyap, G. C., & Bango, M. (2022). Age heaping among individuals in selected South Asian countries: Evidence from Demographic and Health Surveys. *Journal of Biosocial Science, 54*(4), 725–734.

Sironi, M., & Furstenberg, F. F. (2012). Trends in the economic independence of young adults in the United States: 1973–2007. *Population and Development Review*, *38*(4), 609-630.

Sobek, M., Cleveland, L., Flood, S., Ruggles, S., & Schroeder, M. (2011). Big data: Large-scale historical infrastructure from the Minnesota Population Center. Historical Methods, *44*, 61–68.

Sobek, M. (2016). Data prospects: IPUMS-International. In M. White (Ed.), *International handbook of migration and population distribution* (pp. 157–174). Springer.

Speringer, M., Goujon, A., KC, S., Potančoková, M., Reiter, C., Jurasszovich, S., & Eder, J. (2019). *Global reconstruction of educational attainment, 1950 to 2015: Methodology and assessment* (No. 02/2019). Vienna Institute of Demography Working Papers.

Stone, J., Berrington, A., & Falkingham, J. (2011). The changing determinants of UK young adults' living arrangements. *Demographic Research, 25*, 629–666.

Tabutin, D., Schoumaker, B., Coleman, H., Dutreuilh, C., Reeve, P., Tovey, J., & van Hoorn, A. (2020). The demography of Sub-Saharan Africa in the 21st century. *Population, 75*(2), 165–286.

Tasciotti, L., & Wagner, N. (2018). How much should we trust micro-data? A comparison of the socio-demographic profile of Malawian households using census, LSMS and DHS data. *The European Journal of Development Research, 30*(4), 588–612.

UNESCO. (2016). *Creating sustainable futures for all: Global education monitoring report 2016—Gender review.* UNESCO.

UNESCO. (2018). *Global education monitoring report 2019: Migration, displacement and education—Building bridges, not walls.* UNESCO.

UNESCO Institute for Statistics (UIS). (2019). *New methodology shows that 258 million children, adolescents and youth are out of school* (Fact Sheet No. 56). http://uis.unesco.org/sites/default/files/documents/new-methodology-shows-258-million-children-adolescents-and-youth-are-out-school.pdf

UNESCO. (2019). *From access to empowerment: UNESCO strategy for gender equality in and through education 2019–2025.* UNESCO.

UNICEF. (2015). *Key milestones since 1995*. https://mics.unicef.org/key-milestones

UNICEF. (2020). *Multiple Indicator Cluster Surveys (MICS).* http://mics.unicef.org/

United Nations. (2008). *Designing household survey samples: Practical guidelines.* United Nations.

United Nations. (2013). *The Millennium Development Goals report 2013.* United Nations.

United Nations. (2025). *Principles and recommendations for population and housing censuses* (Rev. 4). United Nations.

United Nations. (2026). *Demographic Yearbook System.* https://unstats.un.org/unsd/demographic-social/products/dyb

United Nations Children's Fund. (2015). *Monitoring the situation of children and women for 20 years: The Multiple Indicator Cluster Surveys (MICS) 1995–2015*. UNICEF.

United Nations Children's Fund. (2019). *Every child learns: UNICEF education strategy 2019–2030.* UNICEF.

U.S. Census Bureau. (2019). *Historical Table 1: Percent childless and births in the past year per 1,000 women, CPS, selected years 1976–2018.* https://www.census.gov/data/tables/time-series/demo/fertility/his-cps.html

U.S. Census Bureau. (2020a). *Table MS-2: Estimated median age at first marriage, 1890 to present.* https://www2.census.gov/programs-surveys/demo/tables/families/time-series/marital/ms2.xls

U.S. Census Bureau. (2020b). *Table AD3: Living arrangements of adults 25 to 34 years old, 1967 to present.* https://www2.census.gov/programs-surveys/demo/tables/families/time-series/adults/ad3-25-34.xls

U.S. Census Bureau. (2020c). *Table HH-1: Households by type, 1940 to present.* https://www2.census.gov/programs-surveys/demo/tables/families/time-series/households/hh1.xls

U.S. Census Bureau. (2020e). *Table AD3: Living arrangements of adults 18 to 24 years old, 1967 to present.* https://www2.census.gov/programs-surveys/demo/tables/families/time-series/adults/ad3-18-24.xls

Van Hook, J., & Glick, J. E. (2007). Immigration and living arrangements: Moving beyond economic need versus acculturation. *Demography, 44*(2), 225–249.

Ventura, S. J., & Bachrach, C. (2000). Nonmarital childbearing in the United States, 1940–1999. *National Vital Statistics Reports, 48*(16). National Center for Health Statistics. https://www.cdc.gov/nchs/data/nvsr/nvsr48/nvs48_16.pdf

Verma, V. (1993). *Sampling errors in household surveys.* United Nations Statistical Division.

Verma, V., & Lê, T. (1996). An analysis of sampling errors for the Demographic and Health Surveys. *International Statistical Review, 64*(3), 265–294.

Vikram, K., Vanneman, R., & Desai, S. (2012). Linkages between maternal education and childhood immunization in India. *Social Science & Medicine, 75*(2), 331–339.

Wang, L. (2003). Determinants of child mortality in LDCs: Empirical findings from demographic and health surveys. *Health Policy, 65*(3), 277–299.

Wright, J. A., Yang, H., & Walker, K. (2012). Do international surveys and censuses exhibit “dry season” bias? *Population, Space and Place, 18*(1), 116–126.

Yadav, A., Vishwakarma, M., & Chauhan, S. (2020). The quality of age data: Comparison between two recent Indian censuses 2001–2011. *Clinical Epidemiology and Global Health, 8*(2), 371–376.