Statistical coherence of primary schooling in population census microdata: IPUMS-International integrated samples compared for fifteen African countries

Robert McCaa, Lara Cleveland, Patricia Kelly-Hall, Steven Ruggles, and Matthew Sobek

Corresponding author:

Robert McCaa, rmccaa@umn.edu

Minnesota Population Center, 50 Willey Hall

225 19th Ave. S.

Minneapolis, MN 55455 USA

global cell: 1+952.33.IPUMS (952.334.7867)

fax: 1+612.626.8375

Acknowledgments

Research for this paper was funded in part by the National Science Foundation of the United States, grant SES-0851414 International Integrated Microdata Series (IPUMS-International). The authors gratefully acknowledge the statistical offices that provided the underlying microdata making this research possible: National Institute of Statistics and Demography, Burkina Faso; Central Bureau of Census and Population Studies, Cameroon; Ghana Statistical Services, Ghana; National Statistics Directorate, Guinea; National Bureau of Statistics, Kenya; Institute of Statistics and Geo-Information Systems, Liberia; National Statistical Office, Malawi; National Directorate of Statistics and Informatics, Mali; Department of Statistics, Morocco; National Bureau of Statistics, Nigeria; National Population Commission, Nigeria; National Agency of Statistics and Demography, Senegal; Statistics South Africa, South Africa; Bureau of Statistics, Tanzania; Bureau of Statistics, Uganda; and Central Statistics Office, Zambia. The authors alone are solely responsible for errors.

Abstract

The IPUMS-International project, now in its fifteenth year, integrates and disseminates population microdata for twenty-two African countries (82 countries world-wide) and the number continues to increase as more National Statistical Offices cooperate with the initiative. Statistical quality is a serious concern both for the producers of the microdata as well as the researchers who use them. This paper applies the intra-cohort comparison method to pairs of integrated (harmonized) samples for fifteen African countries to assess statistical coherence using as a benchmark the proportion completing primary school by single years of birth. Samples for six countries show near perfect coherence (R^2 > .9, and regression coefficients ~1.0 +/-<0.08). For a second

group of five countries, coefficients are only slightly larger ($R^2 > 0.6 < 0.9$). Large deviations from 1.0 characterize samples for only four countries. On the whole, the results suggest that samples for the fifteen countries have considerable utility for socio-demographic analysis.

Key words: IPUMS-International, census microdata, metadata, statistical coherence, population census, primary schooling, educational attainment, data quality

Résumé

Le projet IPUMS-International entre dans sa quinzième année, et a pour but de préserver, d'harmoniser (intégrer) et de disséminer les données individuelles des recensements de population. Actuellement il couvre les données de 22 pays africains et de 82 pays à travers le monde. Le nombre de pays couverts augmente chaque année, au fur et à mesure que les instituts nationaux de statistique acceptent de collaborer à cette initiative. L'évaluation de la qualité des données reste une question importante, tant que pour les producteurs que pour les utilisateurs des données individuelles. Cet article utilise la méthode de comparaison intra-cohortes entre des couples de données harmonisées (intégrées) d'échantillons de recensement de 15 pays africains pour évaluer leur cohérence statistique, en prenant l'exemple de la proportion de personnes ayant terminé l'école primaire, classées par année de naissance. Dans un premier groupe de 6 pays, on trouve une cohérence presque parfaite entre les recensements successifs ($R^2 > 0,9$; coefficients de régression b ~1,0 ±0,08). Dans un second groupe de 5 pays, les coefficients sont un peu plus faibles, mais restent acceptables ($R^2 > 0,6$; b<0,9). Les écarts à l'unité plus marqués se trouvent dans les échantillons des quatre derniers pays. Dans l'ensemble, les résultats de cette analyse montrent que ces échantillons sont fiables et peuvent être utilisés pour l'analyse sociodémographique.

Mots-clés: IPUMS-International; Recensement de population; Données individuelles (micro-données); Métadonnées; Cohérence statistique; Scolarisation primaire : Niveau d'instruction ; Qualité des données ; Etude de cohorte.

Introduction

The IPUMS-International project, now in its fifteenth year, disseminates more than 250 integrated census microdata samples representing 82 countries to researchers across Africa and the world. The microdata are disseminated at no cost, but they are not "public." Access is restricted to researchers and policy makers who agree to the stringent conditions-of-use license. Currently, more than 10,000 approved users representing 130 nationalities and countries of residence, may access over 615 million person records representing four-fifths of the world's population free-of-cost thanks to the generous cooperation of National Statistical Offices and Census Agencies world-wide. By 2020, the database is likely to double as the 2010-round of censuses and the backlog of microdata entrusted to the project are integrated into the database. In addition, further expansion is likely as additional NSOs not yet cooperating with the project take the bold step to do so (Table 1).

Table 1 near here

Population census data are collected by nations at great expense and have enormous capacity to inform public policy (Akinyemi and Isiugo-Abanihe. 2014; Reed and Mberu 2014). They are among the most widely used data sources in the social sciences and are broadly employed by policy makers, researchers, journalists, teachers, students, and others. Given the societal investment in censuses and their widely-recognized utility, it is essential that the data be disseminated in a timely manner to maximize their utility.

IPUMS offers a means of disseminating microdata which complements the dissemination activities of National Statistical Offices. NSOs disseminate official statistics and official statistical products to a large number of publics—citizens, officials, the media, analysts, etc. The IPUMS-International project disseminates to a tiny, but important constituency—researchers and policy makers who require detailed data on individuals and households to measure and analyse complex relationships, often making comparisons over time and between nations.

Statistical quality is of great concern to all users of census data, particularly of census microdata. Assessment of data quality makes up an important component of Tools for Demographic Estimation (Moultrie 2014). This paper analyses census quality by comparing statistics from selected pairs of integrated microdata samples disseminated from the IPUMS-International website. We test the statistical coherence of educational attainment, specifically primary schooling completed from two successive censuses for fifteen African countries encompassing one-half (46.7%) of the population of the continent.

Baffour and Valente, in a recent review, define census quality as "fitness for use" and argue that it is characterized by six elements or dimensions: relevance, accuracy, timeliness, accessibility, interpretability, and coherence (2012:122). In this paper we are concerned with coherence, although accuracy and coherence are obviously interrelated.

We experiment with a single dimension of quality, coherence, for a single indicator, primary schooling completed. We ask the question, how do sample statistics for educational attainment from the most recent census compare with a prior census, commonly ten years earlier, for the same country? We test primary schooling completed not only because universal primary education is a Millennium Development Goal but also because it is measured by most African population censuses and widely-available in IPUMS integrated samples.

We use the demographic concept of birth cohort to generate a series of estimates for each individual year of age from 15 to 89 years for each sample. Figures from successive samples are then compared. Where statistics are coherent from one census to the next, they will show the same or closely similar percentages completing primary schooling, birth year-by-birth year.

The results of the experiment are quite promising, indicating a remarkable degree of statistical coherence in percentages of primary schooling completed, as we will show with examples from the 2010 round censuses for Zambia (weighted Pearson product moment correlation coefficient (R^2)=.99, regression coefficient [b]=.97), Burkina Faso (R^2=.98, b=1.03), Kenya (R^2=.97, b=1.02), and for other countries across the continent of Africa. Statistical coherence is increasing over time as we demonstrate by comparisons with the 1990 and earlier rounds of census samples. Nonetheless the range for all fifteen countries is wide (R^2=.38-.99, b=.46-1.37), suggesting significant variations in statistical coherence in successive pairs of census samples.

Soon the experiment may be extended to four additional countries—Ethiopia, Lesotho, Mozambique, and Namibia— once the IPUMS team completes the integration of the corresponding microdata already entrusted to the project. NSOs of ten African countries (Benin, Botswana, Cape Verde, Central African Republic, Cote d'Ivoire, Guinea Bissau, Mauritius, Niger, Rwanda, and Tunisia) have endorsed IPUMS-International protocols but have not yet entrusted the microdata for the 2010 round census. Sixteen others have not yet accepted invitations to cooperate with the IPUMS initiative: Algeria, Angola, Burundi, Chad, Comoros, the Republic of Congo, the Democratic Republic of Congo, Gabon, the Gambia, Libya, Mauritania, Nigeria—National Population Commission—Somalia, Swaziland, Togo, and Zimbabwe.

Data and Methods

IPUMS integrated microdata and metadata. The principal benefit of IPUMS-International to researchers and National Statistics Offices alike is the integration of several decades of population microdata samples for each country—typically beginning with the earliest census for which microdata exist or are recoverable and continuing through the 2010 round. When the project began 15 years ago, few NSOs disseminated census microdata. Today most do. Nonetheless, even today, few NSOs write new documentation to facilitate comparative analysis of two or more censuses. Even fewer re-examine earlier censuses to produce correspondence (cross-walk) tables to harmonize variables in successive samples. Most statistical offices are severely under-staffed and face significant financial and human resource constraints. The general practice

among NSOs is to simply draw a sample, anonymize it, and release it along with whatever ancillary documentation that may be readily available. Often, little guidance is offered to data users on how to compare microdata from successive censuses much less to make international comparisons. Instead, each individual researcher is left with the task of puzzling through the metadata to attain some sort of comparability. Most researchers faced with such a daunting undertaking simply resort to analyzing the most recent census and ignore earlier ones.

IPUMS-International empowers researchers to analyze multiple census years and even multiple countries as a single pooled set of data, facilitating comparative analysis over time and space. High-precision census samples are integrated, variable-by-variable, code-by-code, using a composite coding scheme (Esteve and Sobek, 2003). Integrated metadata are written based on the meticulous study of comprehensive original source documentation. The integration process is strengthened by extensive analysis of the microdata. Thousands of hours are devoted to analyzing, discussing, debating, testing and re-testing by the IPUMS team until the microdata integration is validated for dissemination to researchers. Each year, the process must be repeated as additional census samples are integrated into the database.

The basic goal of the harmonization effort is to simplify the use of the microdata while losing no meaningful information. This is a challenging task because to make data simple for comparative analysis across time and space, it is necessary to develop comparable coding schemes that can be applied to all census samples. Microdata are integrated so that identical concepts (variables, categories) are assigned identical codes. To avoid the loss of important information for those samples that have even more detail, a composite coding strategy is used to retain all original detail, and at the same time provide comparable codes across samples. With composite codes, researchers may easily compare across time and space, yet nuances in meaning are also readily understood.

The first digit, called the "general code," provides information that is available across all samples (the lowest common denominator). The next one or two digits provide additional information available in a substantial subset of samples. Trailing digits provide details that are only rarely available. Where information is not available at the level of a particular digit, a zero place-holder is assigned.

For our analysis of coherence, we focus on the IPUMS-International educational attainment variable, "EDATTAN," the most widely-used variable in the database. Most census microdata with information on this measure indicate four levels or stages, following the ISCED scheme: (1) whether the respondent completed no level of schooling at all, or completed some level: (2) primary, (3) secondary or (4) tertiary or higher schooling. Thus, the first digit of the IPUMS-International composite code consists of four categories (1-4), plus codes for missing data (9) and "not in universe" (0—for children too young to attend school or for others to whom the question was not addressed).

Many census samples contain further information indicating, for example, those who attended primary, secondary or even tertiary schooling, but did not complete the course of study. The second digit captures this information. The third digit distinguishes between technical and general or other tracks. Successful international integration must document such distinctions so that researchers may readily be informed of these and thousands of other details.

Table 2 near here

Table 2 illustrates the general and detailed coding schemes for the educational attainment variable for 20 African countries (represented by the two-digit ISO 3166 country code). Five countries are represented in this table but not in the analysis because the 1990 round sample does not offer the education variables needed to construct EDATTAN (Egypt, Rwanda) or the microdata do not exist or are not presently integrated into the IPUMS database (Sierra Leone, South Sudan, and Sudan). Hopefully, for the next census round the education module in the censuses of these countries will continue to be closely aligned with international standards.

As the upper section of Table 2 shows, all samples have each of the four general levels: less than primary completed, and primary, secondary and tertiary completed. In the lower section of the table, the array of detailed codes displays the considerable variability from country-to-country regarding the levels of schooling completed.

The frequencies in each cell refer to the simple, un-weighted case counts for the corresponding code and sample. The counts are wholly descriptive. As we shall see in the next section, assessment of coherence can be ascertained by using weighted percentages of the codes cross-tabulated by year of birth.

The IPUMS-International team writes metadata for six types of information for each integrated variable in the database. These six and links to the original source documentation are available via seven "tabs" on the IPUMS-International variables pages (see Figure 2 below):

1. Codes
2. General descriptions
3. Comparability discussions
4. Statements of universe
5. Availability of concepts
6. Detailed wording of the original texts ("Questionnaire text," which in turn links to the original source metadata in the official language and English translation) and
7. Links to the "source variables" used in constructing each integrated variable.

The goal of IPUMS integrated metadata is to facilitate informed analysis of the microdata by providing as much essential information as feasible—all readily accessible from the website by means of a few clicks. Note that the metadata are open access. Only the microdata must be restricted

to respect the conditions of use agreed to by the participating statistical agencies.

Definitions: census coherence and primary schooling completed. The Sixteenth Meeting of the United Nations Economic Commission for Europe Group of Experts on Population and Housing Censuses defines coherence as follows (see UNECE 2014, p. 4, Section B.4.f):

> Coherence reflects the degree to which census information can be successfully brought together with other statistical information within a broad analytical framework and over time. The use of standard concepts, definitions, and classifications—possibly agreed at the international level—promotes coherence.

Baffour and Valente (2012:126) identify two types of coherence: internal (results for a single census are coherent within themselves) and external (comparisons between two or more censuses or other official sources). To achieve statistical coherence, definitions, concepts, frameworks and classifications must be clear and consistent both nationally and internationally. When these change, explanations are essential to describe similarities and differences between the old and the new. Baffour and Valente conclude that "ideally the [census] questions should keep the historical formulation to facilitate longitudinal comparison," and any unusual trends or inconsistencies in the data should be explained.

For the 2010-round of censuses, the United Nations Statistics Division recommended "educational attainment" as a core topic and, in post-census processing, recommended the use of categories of the 1997 revision of the International Standard Classification of Education (ISCED) to facilitate international comparisons (UNSD 2008:149-150). "ISCED 1" constitutes primary education, typically 4-7 years completed with six years the most common (UNESCO 2012:17).

The intra-cohort comparison method. A population census contains within it the demographic history of a nation and its people. Successive, high quality censuses of a nation should tell similar, coherent stories. The population historian's tool kit—the principal investigators of IPUMS-International are historians—includes the intra-cohort comparison method, in which a statistic is measured by birth cohorts in successive censuses.

For external coherence we ask the simple question: For each birth year, is the proportion completing primary school in a 2010-round census similar to that for the 2000- or earlier round? Using Kenya as an example, we pose the question: Is the proportion completing primary school of those born in, say, 1965 the same in the sample for the 2009 census as for 1999? As a matter of fact, the answer is yes, almost exactly: 71.4% of those born in 1965 completed primary schooling according to the 2009 census sample compared with 71.3% in the 1999 sample—a minute difference and a remarkable testimony to statistical coherence!

We extend the question to encompass an entire series of birth cohorts, beginning 15 years before the census (very few individuals complete primary

school at a more advanced age) and extending back in time until the absolute frequencies become too small to be reliable, say beyond age 89. We use the product moment correlation and the least-squares regression coefficients to measure the degree of coherence for each pair of series. The older census of the pair is used to predict the percentage for the more recent census. The sample counts of the first census are used as weights. For Kenya 1999 compared with 2009, we find R^2=.97 and b=1.02, which indicates an exceedingly high degree of coherence although not a perfect 1.0 (see Figure 1 and Table 5 below). In addition we must take into account sample error. The 95% confidence interval for our estimate of the regression coefficient is +/-.05, that is the range is .97-1.07, indicating an outstanding degree of statistical coherence.

There are at least three caveats for assessing external coherence: census agency practices, IPUMS harmonization, and bias. First, the questions, definitions and categories posed in successive censuses and the training of the field enumerators must be taken into account as well as how the data were processed and edited by the national census authority. Second, since we are analysing IPUMS integrated samples, we must also consider how the IPUMS team harmonized the microdata, and whether the decisions taken to integrate coding schemes in successive censuses are correct or not. Third, the method assumes that there are no differentials in mortality, migration or reporting by level of educational attainment. Where the less educated suffer from higher mortality rates then this will introduce a systematic upward bias to the proportions completing primary education. Likewise where the likelihood of migration to or from another country is associated with educational attainment, then lack of coherence will be exaggerated by international movements unrelated to the quality of census operations. Then, too, there may be bias in reporting by the respondents, particularly where educational attainment is low and ages are reported in rounded approximations, such as 40, 50, 60 or 65, 75, 85, etc. The illiterate are particularly prone to misstating age. For additional details of the method see Feeney (2014).

In addition, note that our analysis is confined to sample microdata. Using the full-count microdata would eliminate sample error and would enhance the analysis below. The reader is cautioned that these are unofficial estimates based on samples. Moreover, our estimates of standard errors are not robust because we do not take into account clustering, stratification or differential weighting that may be inherent in the sample designs for specific censuses.[1]

Results

Kenya. Figure 1 depicts the primary school completion rates by year of birth, as computed from the IPUMS integrated samples for the 1979, 1989, 1999 and 2009 censuses of Kenya. The curves reveal astonishing coherence, with product moment correlation coefficients of .97-.99 for comparisons of 1999/2009, 1989/1999 and 1979/1989. Regression coefficients (aside from 1979:1989) of 0.99 and 1.02 are amazingly close to 1.0, which would indicate perfect coherence. Perhaps there should be little surprise that the results are so

nearly identical because all sets of data were produced by a single statistical agency. Nevertheless, the underlying data in each case were collected by 50-100 thousand field workers conducting face-to-face interviews at four different points in time, separated by intervals of ten years each. The data were processed and coded using increasingly sophisticated technologies that nonetheless offer many opportunities for error. Furthermore, the figures are computed from samples using integrated variables constructed by the IPUMS team with no knowledge that the microdata might be examined this way. Nonetheless the statistical coherence of the results in Figure 1 is striking. Researchers should take comfort in the outstanding coherence between successive census samples of Kenya.

Figure 1 near here

Consider, too, the three caveats referenced in section four: census agency practices, IPUMS harmonization, and bias.

First, as can be seen in Table 3, the KNBS designed questions on educational attainment that, on the whole, are quite consistent from census-to-census—even with closely similar lay-outs. Since 1989, the Kenyan censuses request the level of schooling completed. Despite the fact that the 1979 census refers to highest level reached, figures seem to be understated relative to later censuses. Regarding attendance, the 2009 and 1989 forms offered four options—at school, left school, never went to school, NS/DK (not-stated/do-not-know). The 1999 form added an under-five-years category while the 1979 dropped the NS/DK category.

Table 3 near here

Second, the IPUMS harmonization of educational attainment offers two variants: an international and a national recode, EDATTAN, and, in the case of Kenya, EDUCKE, respectively. For the international variable, only nine codes are needed compared with 24 for EDUCKE. The IPUMS comparability discussion consists of 250 words for EDATTAN versus 400 for EDUCKE. The EDATTAN comparability text begins as follows:

> Kenya changed its educational system in 1985 to an 8-4-4 system (8 years of primary education, 4 years of secondary, and 4 years of university). Previously, Kenya had used a 7-6-3 system. In the Kenyan censuses, respondents simply provided the highest grade level completed (see EDUCKE) without any reference to a specific education system structure. Moreover, responses in all Kenya samples include standard 1 to 8 (primary) and forms 1 to 6 (secondary), which do not exactly match either education system.
>
> For EDATTAN, Kenya is coded into a 6-3-3 structure: standard 6 (or more) is interpreted as completion of primary, form 1 (or more) as completion of lower secondary, and form 4 (or more) as completion of upper secondary. This may overestimate educational attainment for some persons who only had 7 (and not 8) years of primary education.[2]

As the quotation indicates, for the international recode we impose where possible a standard of six years of education across all samples in the database. Note that the IPUMS system offers researchers the opportunity to easily construct recodes using their own criteria or to deconstruct IPUMS codes to check consistency and accuracy of the countless decisions made in the harmonization process. The "Source Variable" tab in Figure 3 points the way. For our purposes this figure explains that primary schooling completed was coded from the number of years of schooling question in each census with a code of six years or more required to satisfy the condition. Note that we applied to the extent possible this standard throughout the entire database, even though, in the case of Kenya, since 1985 the National Education System defines primary schooling as completed with eight years of attendance.

Researchers studying only a single country may favor (and download) the national recode variable, such as EDUCKE, while others interested in comparing differences between countries are likely to favour the international variant, EDATTAN.

Figure 2 near here

In assessing external coherence, the third caveat—bias introduced by the assumptions regarding migration, mortality, and reporting— should also be considered. The fact that the 1979 proportions are systematically lower at every age than 1989 may suggest that the less educated have slightly worse survival chances or higher out-migration rates than the better educated. An upward bias in reporting events more distant in the past is to be expected, although in the present case the bias would seem to be nil, because, subtracting the percentage for each birth year for 1999 from 2009, we find the mean difference is a miniscule 0.2 percentage points and the median is 0.0. We conclude that the coherence of Kenyan censuses with respect to primary schooling completed is exceedingly high, comparable to coherence rankings for developed countries (based on secondary completed), despite the fact that primary education is not yet universal even for the youngest generation of Kenyans.

Nigeria. For a second test, we turn to Nigeria, where the microdata are not from population censuses but instead are nationally representative samples, General Household Surveys (GHS), conducted by the National Bureau of Statistics. While the GHS is conducted annually and all are disseminated by IPUMS-International, our initial discussion focusses on only two: 2006/7 (April-March) and 2010/11 (July-March).

GHS Instructions to field workers are relatively un-changing, but the form has changed significantly and may be the source of some confusion not only for interviewees but also for field workers and data processing personnel. To examine this issue in detail, from www.ipums.org/international click "Browse and Select Data," "Select Samples," pick "Nigeria," "Submit Sample Selections," from the Variable selection Person drop-down menu select, "Education," click "EDATTAN," and finally click the "Questionnaire Text" tab on the variables metadata page https://international.ipums.org/international-action/variables/ EDATTAN#questionnaire_text_section. To appreciate the

1

differing text and lay-outs from survey-to-survey, scroll through the 1,557 words which appear. The 2006/7 form has space to code two digits. For 2007/8 – 2009/10 nineteen pre-coded tick boxes are offered. For an individual completing only the first year of primary the field worker ticked the "03" box and for sixth year completed "09". For the 2010/11 survey, the form offers a composite scheme totalling twenty-seven boxes with the first digit indicating level and the second the number of years completed. Thus for primary with six years completed the interviewer ticked box "16".

Overall, a low level of coherence is apparent in the summary statistics: R^2=.38, b=.46, and the mean difference is five percentage points. Indeed, in terms of coherence, the General Household Survey samples are among the lowest in the entire IPUMS-International collection. To discount error in one or another of the samples, we examined the entire series of surveys to construct a matrix of correlations and regression coefficients, only to find that low coherence characterizes the entire lot (Table 4).

Table 4 near here

In the absence of census microdata for Nigeria, we decided to examine unofficial microdata: Post Enumeration Surveys (PES) for the 1991 and 2006 censuses, which the National Population Commission has kindly entrusted to the Minnesota Population Center.[3] We were astonished and pleased to find that the PES are quite coherent with R^2=.83, b=.92. The 1965 year-of-birth benchmark is exactly the same for both samples at 57.8% (Table 5). Mean and median differences are one-half to one-third those of the GHS. The diagonal of Table 4 reports the mean percentage completing primary schooling for the 2 years before and after the 1965 benchmark. We see that the perfect alignment for the year 1965 is a fluke because the average for 1963-1967 is 52.0% for the 1991 enumeration and 56.4% for 2006 or a spread of 4.4 percentage points. For the GHS the spread ballons to 13.2, exactly three times that for the PES (subtract 47.2% from 60.4%, 2006/7 and 2010/11, respectively).

Figure 3 depicts the complete scatter for all birth cohorts of primary school completion percentages for the PES (top panel) and the GHS (bottom panel). Severe digit attraction for ages ending in zero and five cloud the picture, but the greater coherence of the census PES stands out.

Unsatisfied with the low coherence of the GHS, we drilled down into the IPUMS integrated samples and found that in the most recent survey (GHS2010/11) for the first time year and month of birth were recorded on the survey form. Both are integrated into the IPUMS database. By computing age from year of birth and year of sample, our measures of association improve markedly with R^2 increasing from 0.54 to a respectable .84 and the regression coefficient from .66 to 1.14. In other words, the National Bureau of Statistics implemented corrective measures in the 2010/11 General Household Survey that greatly enhance statistical coherence for intra-cohort comparisons.

Figure 3 near here

Table 5 offers additional statistical detail for assessing coherence for pairs of Nigerian samples. Strong digit-preference in age reporting of the uneducated distorts any chronological comparison, as we see for both the PES

and GHS samples. The Whipple age heaping index—we use the "all digit" or total index developed by Spoorenberg and Dutreuilh (2007)—reported in Table 5 indicates that age declarations in the Nigerian samples analyzed are "very rough" with the PES slightly better than the GHS. For GHS2010/11 computing age from birth year as described above greatly improves the Whipple total index from 5.86 to 3.83.

Figure 4 and Table 5 near here

Additional country comparisons. Table 5 and Figure 4 summarize our analysis for African countries with pairs of samples disseminated by IPUMS-International (April 2015). Nearly perfect coherence is attained by six sets of samples—those for Burkina Faso, Kenya, Morocco, South Africa, Tanzania and Zambia. These show a mean difference of less than one percentage point, R^2 => .93, and less than +/-0.08 deviation from unity for regression coefficients. A second group—with mean differences slightly greater and coefficients slightly larger—characterize pairs for Ghana, Malawi, Nigeria PES, and Uganda. A third cluster with coefficients showing substantial departures from one and wide ranging mean differences is made up of pairs of samples for Cameroon, Guinea, Liberia, Nigeria GHS, and Senegal.

Digit attraction for single years of age, which has little to do with the quality or coherence of census design or execution, explains much of the deviation in statistical coherence. The distortion is accentuated when censuses are taken more than ten years apart, as in the case of all pairs of African samples with correlations below .93 analysed in this paper.

Mali. Take the case of Mali, for example. As Table 5 indicates, the three most recent censuses are conducted eleven years apart and the level of digit attraction is high with Whipple total indices exceeding 3.0. Shifting the year of birth back one year for the 1998 census and two years for the 2009 enumeration to synchronize the zero digit for age in the three censuses doubles R^2 to .91, lifts the regression coefficient within .03 of unity and shrinks the 95% confidence interval by two-thirds. The Malian microdata are certainly "fit for use" if number of publications is our yard-stick. The IPUMS-International bibliography yields 18 citations referencing Mali, including three chapters in the recently released Continuity and Change in Sub-Saharan African Demography (Clifford O. Odimegwu and John Kekovole, eds.) and a chapter in World Population and Human Capital in the Twenty-First Century (Wolfgang Lutz, William P. Butz and Samir K.C., eds.).[4]

Discussion.

Coherence in successive censuses, as measured by the intra-cohort comparison method, is a strong statistical test. Nonetheless intra-cohort coherence is rarely assessed because the method is difficult to apply unless the microdata are integrated. However, once integrated into a single database, such as the case with samples disseminated by IPUMS-International, the method is easily applied to variables characterized by a "rite-of-passage," such as graduation from primary, secondary, or tertiary schooling, literacy, children ever-born, ever-married, etc.

Researchers must understand that the IPUMS-International integrations are performed *ex-post-facto.* The National Statistical Agency-owners of the data are not responsible for the decisions taken to design the IPUMS system nor for the resulting integrated codes. In contrast, Eurostat's Census Hub dissemination platform was constructed by European statistical offices before the 2010-round of censuses were taken so that integration of concepts, definitions, and categories was designed into the system prior to the actual taking of the censuses.[5]

The challenge for the IPUMS integration team is to deal with statistical facts as they exist in each individual set of microdata with no input on census definitions or designs. Thanks to the widespread acceptance of United Nations Statistics Division's *Principles and Recommendations for Population and Housing Censuses*, harmonization of census concepts and categories is possible to a greater or lesser degree.

For the 2020-round of censuses statistical coherence is likely to be even greater than for the 2010-round thanks to the ever-increasing cooperation among National Statistics Offices in Africa, the African Centre for Statistics, the United Nations Statistics Division, , and—most of all—the citizens of the countries where the censuses are taken.

References

Akinyemi, Akanni Ibukun and Uchec C. Isiugo-Abanihe. 2014. Demographic dynamics and development in Nigeria: Issues and Perspectives. African Population Studies Vol. 27, 2 Supp (Mar):239-248.

Baffou, B. and P. Valente. 2012. An evaluation of census quality. *Statistical Journal of the IAOS* 28:121-135. DOI 10.3233/SJI-2012-0752.

Esteve, A. and M. Sobek. 2003. Challenges and methods of international census harmonization . *Historical Methods* 36: 66-79.

Feeney, G. 2014. Literacy and Gender: Development Success Stories. *Population and Development Revie*w *40*:545–552. DOI 10.1111/j.1728-4457.2014.00697.

Lutz, Wolfgang, William P. Butz and Samir K.C. (eds.) 2014. World Population and Human Capital in the Twenty-First Century. Oxford: Oxford University Press.

McCaa, R. 2013. The Big Data Revolution: IPUMS-International. Trans-Border Access to Decades of Census Microdata Samples for Three-fourths of the World and more. *Revista de Demografía Histórica 30*: 69-87.

McCaa, R. and A. Esteve. 2006. IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted access census microdata extracts to academic users. *Monographs of official statistics: Work session on statistical data confidentiality*. Luxembourg: Office for Official Publications of the European Communities, 37-46.

Minnesota Population Center. 2014. *Integrated Public Use Microdata Series, International: Version 6.3* [Machine-readable database]. Minneapolis: University of Minnesota.

Moultrie TA. 2013. "General assessment of age and sex data". In TA Moultrie, RE Dorrington, AG Hill, K Hill, IM Timæus and B Zaba (eds). Tools for Demographic Estimation. Paris: International Union for the Scientific Study of Population. http://demographicestimation.iussp.org/content/general-assessment-age-and-sex-data. Accessed 23/03/2015.

Obono, Oka and Elizabeth Omolaubi. 2014. Technical and political aspects of the 2006 Nigerian population and housing census. African Population Studies vol. 27, 2 Supp (Mar): 249-262.

Odimegwu, Clifford and John Kekovole (eds.) 2014. Continuity and Change in Sub-Saharan African Demography. New York: Routledge.

Reed, Holly E. and Blessing U. Mberu. 2014. Capitalizing on Nigeria's demographic dividend: reaping the benefits and diminishing the burdens. African Population Studies vol. 27, 2 Supp (Mar):319-330.

Ruggles, S. 2006. The Minnesota Population Center data integration projects: Challenges of harmonizing census microdata across time and place. *Proceedings of the American Statistical Association, Government*

Statistics Section. Alexandria, VA: American Statistical Association, 1405-1415.

Spoorenberg, T and C. Dutreuilh. 2007. Quality of Age Reporting: Extension and Application of Modified Whipple Index. *Population* (English Edition), 62(4):729-741.

United Nations European Economic Commission (UNECE). 2014. Group of Experts on Population and Housing Censuses. Quality Management – Draft Text. *Conference of European Statisticians Recommendations for the 2020 Census Round*. Geneva.

United Nations Department of Economic and Social Affairs, Statistics Division (UNSD). 2008. *Principles and Recommendations for Population and Housing Censuses, Revision 2*. Statistical papers Series M. No. 67/Revision 2, New York.

UNESCO Institute for Statistics. 2012. *International Standard Classification for Education ISCED 2011*. Montreal.

6

Table 1. IPUMS-International: 2010 census round microdata by year and status for 176 countries and territories

Census year	IPUMS-International Partners		C. Not yet partners (>250,000 population)
	A. 2010 microdata entrusted **disseminating (bold)**	B. 2000 or earlier microdata entrusted; 2010, not yet	
2005-9	**2005 Cameroon, Colombia, Nicaragua, Nigeria (NSSO); 2006 Burkina Faso, Egypt, France, Iran, Ireland**, Lesotho; **2007 El Salvador, Fiji Islands, Palestine, Peru**, **Ethiopia, Mozambique; 2008 Cambodia,** Israel, **Liberia, Malawi, South Sudan, Sudan; 2009 Kenya, Kyrgyz Republic, Mali**, **Kenya**	**2009** Belarus, Guinea Bissau	**2005** Bhutan, Kuwait, Laos, United Arab Emirates; **2006** Hong Kong SAR, Libya, Macau SAR, Maldives; **2007** Congo Republic, French Polynesia, Swaziland; **2008** Algeria, Burundi, Korea DPR; **2009** Azerbaijan, Chad, Djibouti, Kazakhstan, New Caledonia, Solomon Islands
2010	**Argentina, Brazil, Dominican Republic**, **Ecuador, Ghana, India (NSSO), Indonesia, Mexico, Panama, Puerto Rico**, Trinidad and Tobago, **USA (ACS), Zambia**	Cape Verde, China, Korea RO, Malaysia, Mongolia, Philippines, Saint Lucia, Switzerland, Thailand	Bahamas, Barbados, Belize, Finland, Japan, Qatar, Russian Federation, Saudi Arabia, Singapore, Taiwan, Tajikistan, Timor Leste, Togo
2011	**Austria, Armenia, Bangladesh**, Czech Republic, **France,** Greece, Hungary, Iran, **Ireland,** Namibia, **Nigeria (NBS)**, Poland, **Portugal,** Romania, **South Africa, Spain, Uruguay**	Botswana, Bulgaria, Canada, Costa Rica, Germany, Italy, Jamaica, Mauritius, Nepal, Netherlands, Papua New Guinea, Slovak Republic, Slovenia, Turkey, United Kingdom, Venezuela	Albania, Australia, Bahrain, Belgium, Brunei, Croatia, Cyprus, Denmark, Eritrea, Estonia, Iceland, Nigeria (ORG), Latvia, Lithuania, Luxembourg, Malta, Montenegro FYR, Norway, Sweden
2012+		**2012** Bolivia, Chile, Cuba, Paraguay, Rwanda, Tanzania, Turkmenistan; **2013** Benin, Guinea-Conakry, Honduras, Niger, Senegal **2014+** Central African Republic, Cote d'Ivoire, Guatemala, Haiti, Jordan, Madagascar, Morocco, Pakistan, Sierra Leone, Tunisia, Uganda	**2012** Georgia, Guyana, Macedonia FYR, Nauru, New Zealand, Sri Lanka, Suriname, Tuvalu, Zimbabwe; **2013** Bosnia-Herzegovina, Comoros, Gabon, Gambia, Mauritania, São Tome y Principe; **2014+** Angola, Congo DR, Ecuatorial Guinea, Moldova Republic, Myanmar, Somalia

Source: http://www.hist.umn.edu/~rmccaa/IPUMSI/census_microdata_inventory.htm
Census dates: http://unstats.un.org/unsd/demographic/sources/census/censusdates.htm

7

Table 2. EDATTAN (educational attainment): IPUMS-International general and detailed harmonized codes for 20 African countries

Cell counts refer to un-weighted frequencies for the corresponding codes for the most recent sample integrated

	Country (ISO 3166)	BF	CM	EG	GH	GN	KE	LR	MW	ML	MA	NG	RW	SN	SL	ZA	SS	SD	TZ	UG	ZM
Code	Label Census year	2006	2005	2006	2010	1996	2009	2008	2008	2009	2004	2010/11	2002	2002	2004	2001	2008	2008	2002	2002	2010
General																					
0	NIU (not in universe)	149,884	188,657	1,538,130	203,395	169,008	348,786	53,540	179,294	153,528	·	8,679	194,111	82,671	91,883	371,971	·	940,736	609,010	454,611	293,345
1	Less than primary completed	1,118,587	848,125	2,446,662	1,249,174	490,767	1,836,847	211,048	917,091	1,126,404	1,098,721	31,177	519,196	738,152	316,149	1,352,477	510,886	3,412,071	1,825,040	1,406,182	531,993
2	Primary completed	104,472	582,058	1,115,260	756,794	49,394	1,084,063	56,786	175,273	129,869	269,935	15,632	115,283	144,311	74,276	1,351,155	25,278	384,925	1,167,209	558,048	377,811
3	Secondary completed	15,750	89,032	1,627,179	232,457	9,525	449,164	23,882	61,377	18,330	87,648	11,007	9,941	24,026	3,557	588,002	4,309	22,763	119,903	69,108	113,676
4	University completed	6,097	22,985	550,621	24,469	3,487	32,621	2,801	3,007	11,147	26,412	1,357	1,069	5,402	3,155	62,050	1,860	82,312	11,572	9,500	5,148
9	Unknown	23,034	41,502	4,582	·	6,890	90,454	·	5,935	12,578	4	4,339	3,792	·	5,278	·	432	223,723	1	·	·
Detailed																					
0	NIU (not in universe)	149,884	188,657	1,538,130	203,395	169,008	348,786	53,540	179,294	153,528	·	8,679	194,111	82,671	91,883	371,971	·	940,736	609,010	454,611	293,345
100	LESS THAN PRIMARY COMPLETED			2,446,662	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
110	No schooling	956,723	562,779	·	779,326	414,479	962,300	153,528	354,277	886,979	753,293	21,272	211,188	605,788	216,035	533,996	443,237	2,661,467	1,016,026	576,649	245,676
120	Some primary	161,864	285,346	·	469,848	76,288	874,547	57,520	562,814	239,425	345,428	9,905	308,008	132,364	100,114	818,481	67,649	750,604	809,014	829,533	286,317
130	Primary (4 years)	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
	PRIMARY COMPLETED, LESS THAN SECONDARY																				
	Primary completed																				
211	Primary (5 years)	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
212	Primary (6 years)	64,051	440,989	562,409	253,480	32,970	833,964	33,674	114,533	83,452	221,906	10,541	93,132	109,871	47,264	688,090	15,259	203,929	1,108,055	398,834	244,358
	Lower secondary completed																				
221	General and unspecified track	40,421	112,047	552,851	503,314	16,424	250,099	23,112	60,740	46,417	48,029	5,091	15,081	34,440	20,793	663,065	10,019	180,996	59,154	157,207	133,453
222	Technical track	·	29,022	·	·	·	·	·	·	·	·	·	7,070	·	6,219	·	·	·	·	2,007	·
	SECONDARY COMPLETED																				
	General or unspecified track																				
311	General track completed	10,137	30,548	1,481,464	115,331	1,945	333,473	17,373	59,180	2,437	54,995	8,614	2,852	13,260	·	588,002	1,429	9,198	109,399	29,464	63,146
312	Some college/university	5,613	51,852	·	15,429	2,724	13,627	4,797	2,197	7,960	32,653	·	2,331	10,766	·	·	·	·	10,504	3,761	1,483
320	Technical track	·	·	·	·	·	·	·	·	·	·	·	4,376	·	·	·	·	·	·	·	·
321	Secondary technical degree		5,661	·	32,555	2,227	10,966	·	·	3,912	·	297	382	·	·	·	·	·	·	·	·
322	Post-secondary technical education		971	145,715	69,142	2,629	91,098	1,712	·	4,021	·	2,096	·	·	3,557	·	2,880	13,565	·	35,883	49,047
400	UNIVERSITY COMPLETED	6,097	22,985	550,621	24,469	3,487	32,621	2,801	3,007	11,147	26,412	1,357	1,069	5,402	3,155	62,050	1,860	82,312	11,572	9,500	5,148
999	UNKNOWN/MISSING	23,034	41,502	4,582	·	6,890	90,454	·	5,935	12,578	4	4,339	3,792	·	5,278	·	432	223,723	1	·	·

Source: https://international.ipums.org/international-action/variables/EDATTAN#codes_section

Note: "NG 2010/11" refers to the General Household Survey of the National Bureau of Statistics of Nigeria; all others refer to national census samples.

Table 3. Educational attainment questions in the 1979, 1989, 1999 and 2009 population censuses of Kenya differ in details but are generally quite comparable

2009

D: Information Regarding Persons Aged 3 Years and Above

(P-39)	(P-40)	(P-41)
Education		
What is the school/ Learning institution attendance status of <NAME>? 1=At school/ Learning Institution 2=Left School/ Learning Institution 3=Never went to school/ Learning Institution 9=DK	What is the highest Std/Form/Level reached by <NAME>? The code list provided Write "97" if P-39 equals 3 or 9	What is the highest Std/Form/Level completed by <NAME>? The code list provided Write "97" if P-39 equals 3 or 9

1999

B: Information Regarding Persons aged 5 y

Serial No.	Education (P-22)	(P-23)	(P-24)
	What is the school attendance status of.....? 1- At school 2- Left school 3- Never went to school 4- NS / DK 0- Under 5 years	What is the highest level reached by.....? (Write the appropriate code using the categories shown inside the front cover) (code 97 if P-22=3)	What is the highest level completed by .....? (Write the appropriate code using the categories shown inside the front cover) (code 97 if P-22=3)
1			

1989

B. PERSONS AGED 6 YEARS AND OVER

LITERACY	EDUCATION	
Does.... know how to read and write a simple statement in any language ? 0. NA 1. Yes 2. No	Has ..... ever attended school ? 1 At school 2 Left school 3 Never went to school (Code 0 if age is 5 years or less)	What is .....'s highest level of education completed? eg. class, form, university (Write the appropriate code using the categories shown inside front cover)
P 19	P 20	P 21

1979

EDUCATION	
School attenda— nce, State whether AT LEFT NEVER	What was / is the highest class or form reached ?
h	i

Source: https://international.ipums.org/international/enum_materials.shtml
Note: All "enum_materials" on the above link are available in the official language and English (unofficial translation by IPUMS, as necessary).

Table 4: Nigeria: Post Enumeration Surveys compared with General Household Surveys

		R^2 (above the diagonal)						
	% 1963-7	PES1991	PES2006	GHS2006	GHS2007	GHS2008	GHS2009	GHS2010
	PES1991	52.0	.84	.79	.47	.57	.55	.36
	PES2006	.92	56.4	.87	.59	.67	.68	.52
	GHS2006	.92	1.15	47.2	.59	.70	.67	.58
B	GHS2007	.82	.97	.79	58.8	.66	.70	.64
	GHS2008	.81	.99	.85	.82	55.0	.62	.67
	GHS2009	.78	.92	.76	.80	.76	59.2	.64
	GHS2010	.69	.85	.75	.79	.82	.85	60.3

Source: www.ipums.org/international except PES1991 and PES2006 which are unofficial microdata kindly entrusted by the National Population Commission but not yet integrated into IPUMS.

Note: R^2 (product moment correlation coefficient) is displayed above the diagonal; regression (b (regression coefficient), below. The diagonal reports the percentage completing primary school for the 1963-1967 birth cohort. Computations by the authors.

Comment: Comparing Nigeria's census Post Enumeration Surveys And General Household Surveys with respect to R^2 and b shows the PES pair to be more coherent than all GHS combinations despite the fact that the censuses were taken fifteen years apart while for GHS pairs the average difference is two or three years and the maximum is five.

Table 5. Statistical Coherence in Primary Schooling For Pairs of Samples: 15 African Countries

			Harmonized Education Variables			Primary Schooling Completed (EDATTAN)					
		Whipple	SCHOOL	YRSCHL	EDATTAN	Born	~55 Over-lapping Birth Years				
		Index	Attendance	Years	Levels	1965	Mean	Difference		R2	b
Country	Year	(Total)	Number of Categories			%	%	Mean	Median		95% ci
Burkina Faso	2006	2.13	5	-	9	11.1	7.4	0.1	0.1	.98	1.03
	1996	2.74	-	-	8	12.1	7.3				+/-.04
Cameroon	2005	3.16	4	25	12	44.4	40.2	11.8	11.5	.64	.68
	1987	3.08	-	26	12	59.5	28.3				+/-.14
Ghana	2010	2.62	4	20	10	47.2	51.9	5.0	5.5	.93	1.02
	2000	3.56	4	20	10	42.9	46.9				+/-.07
Guinea (Conakry)	1996	4.43	4	25	11	24.5	10.5	1.0	0.9	.43	.65
	1983	4.79	4	25	10	19.1	9.5				+/-.20
Kenya	2009	2.11	5	19	11	71.4	46.3	0.2	0.0	.97	1.02
	1999	2.25	4	16	8	71.3	46.1				+/-.05
Liberia	2008	2.09	4	20	9	40.2	18.5	4.5	3.3	.73	1.37
	1974	3.34	3	19	8	27.1	14.0				+/-.27
Malawi	2008	1.39	4	23	9	29.8	21.3	1.3	0.7	.90	1.19
	1998	2.45	3	20	8	28.7	20.0				+/-.10
Mali	2009	3.49	5	26	11	15.3	9.5	1.2	1.2	.45	.71
	1998	3.57	3	27	12	13.3	8.2				+/-.20
Morocco	2004	1.24	-	22	8	29.4	16.6	0.0	0.3	.99	.92
	1994	1.35	-	20	8	30.1	16.7				+/-.02
Nigeria PES (unofficial,	2006	5.69	6	-	9	57.8	42.2	2.6	1.4	.83	.92
unedited, unweighted)	1991	7.70	-	-	7	57.8	39.7				+/-.11
Nigeria (GHS)	2010/11	5.86	5	19	10	48.0	51.4	5.9	4.4	.38	.46
	2006/07	5.33	3	16	8	47.2	45.6				+/-.14
Senegal	2002	3.63	3	20	8	26.3	16.8	2.7	3.3	.39	.67
	1998	2.11	4	22	9	28.5	14.1				+/-.22
South Africa	2001	0.66	2	15	7	74.2	60.6	0.3	0.1	.99	1.01
	1996	0.69	4	17	8	74.1	60.3				+/-.02
Tanzania	2002	2.83	5	18	9	73.4	30.2	0.5	0.7	.93	.92
	1988	3.84	5	18	9	77.5	29.7				+/-.06
Uganda	2002	1.80	4	16	10	46.1	36.8	4.7	3.6	.89	.89
	1991	2.95	4	21	10	44.2	31.9				+/-.08
Zambia	2010	1.56	4	15	9	61.9	49.0	3.7	3.2	.99	.97
	2000	1.71	4	15	9	57.3	45.4				+/-.04

Source: see table 4. Note: Whipple total index: 0 = best--no digit preference--to 16 = worst--preference for a single digit such as zero; see Spoorenberg and Dutreuilh (2007)

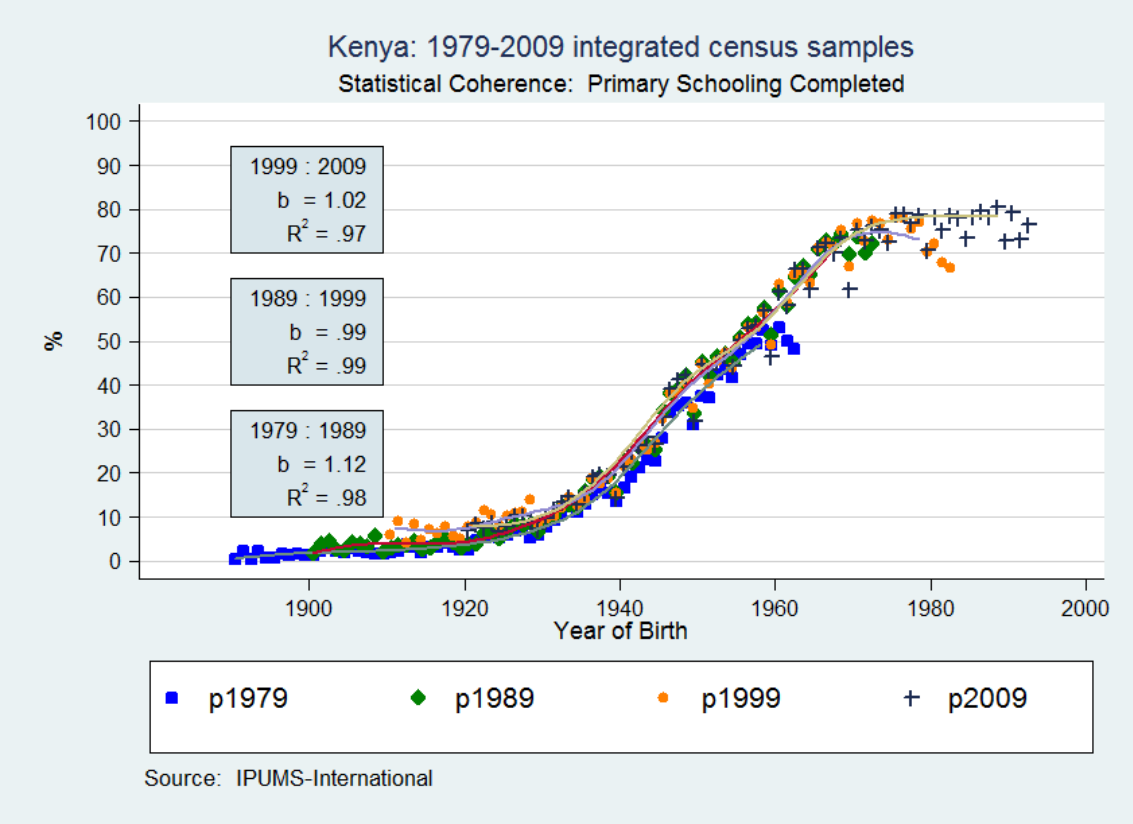
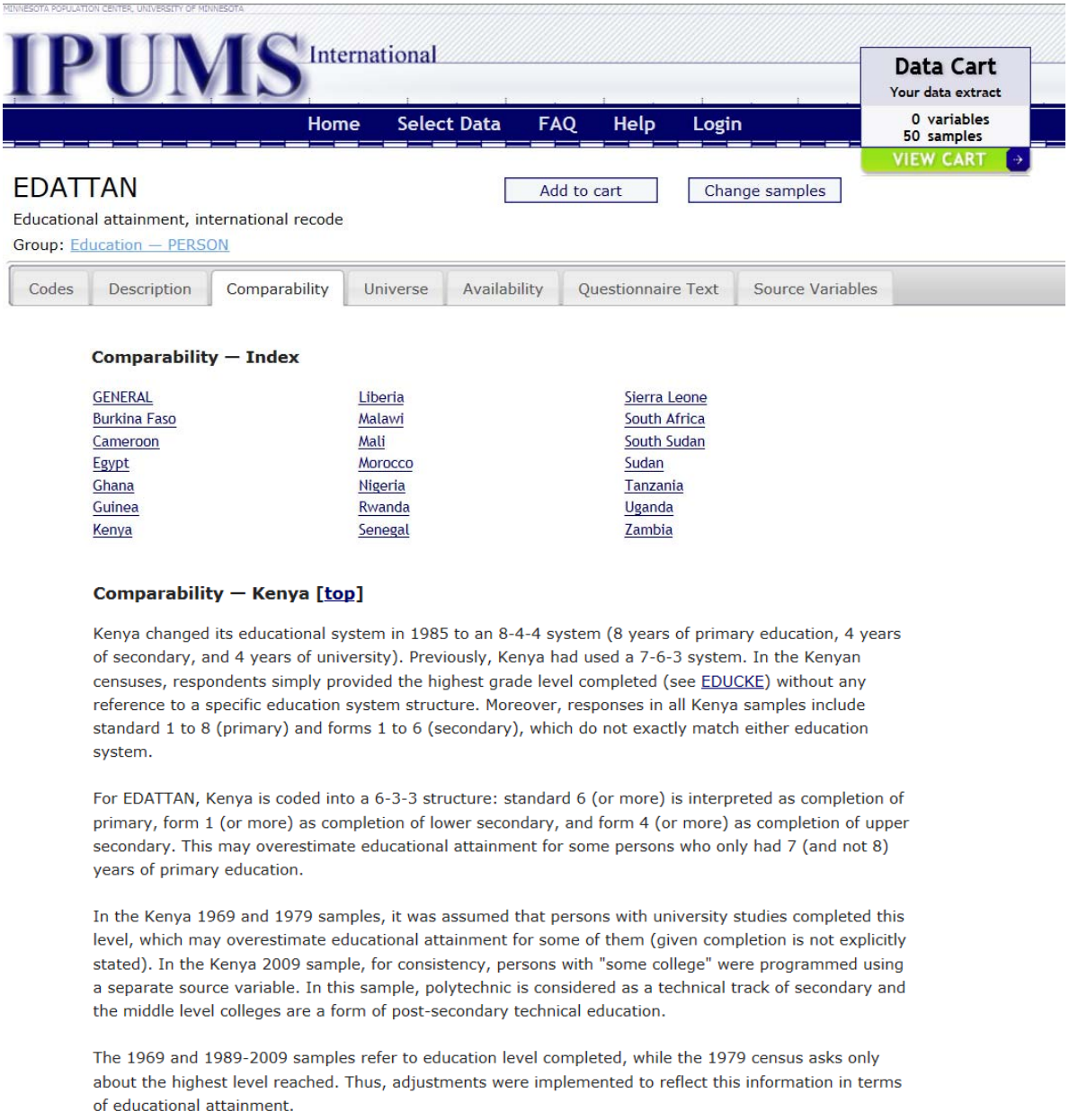
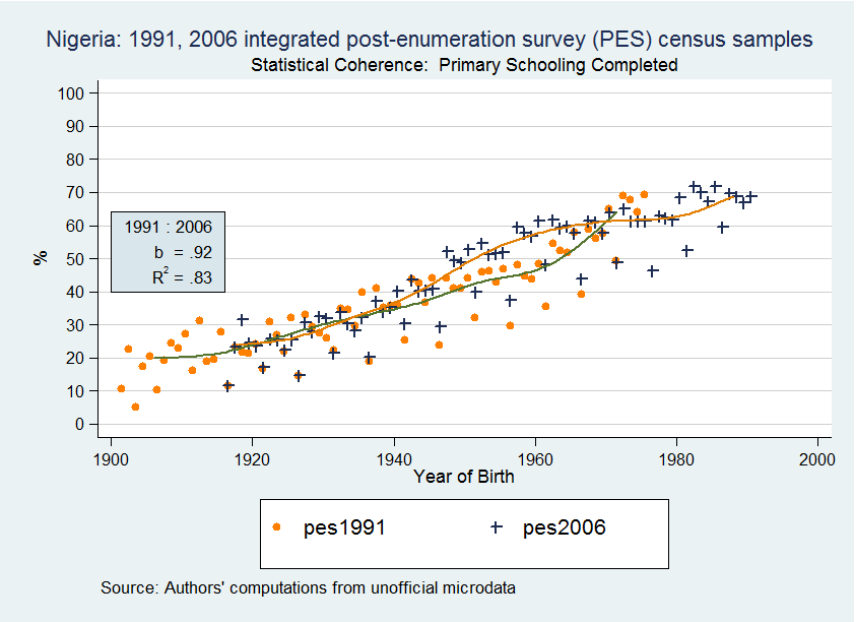


Figure 1. Kenya. Four census samples compared: 2009, 1999, 1989, and 1979 statistical coherence in EDATTAN primary schooling completed



Source: https://international.ipums.org/international-action/variables/EDATTAN#comparability_section

Figure 2. IPUMS-International metadata for educational attainment – international recode comparability discussion: Kenya



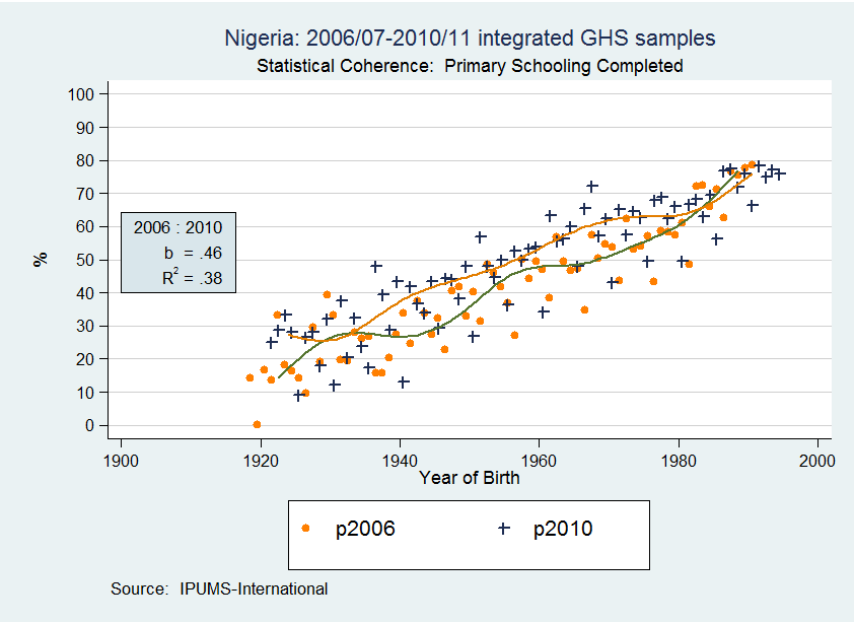


Figure 3. Nigeria. Census Post Enumeration and General Household Surveys Compared

Figure 4. Statistical coherence of primary schooling completed in IPUMS integrated census samples of twelve African countries

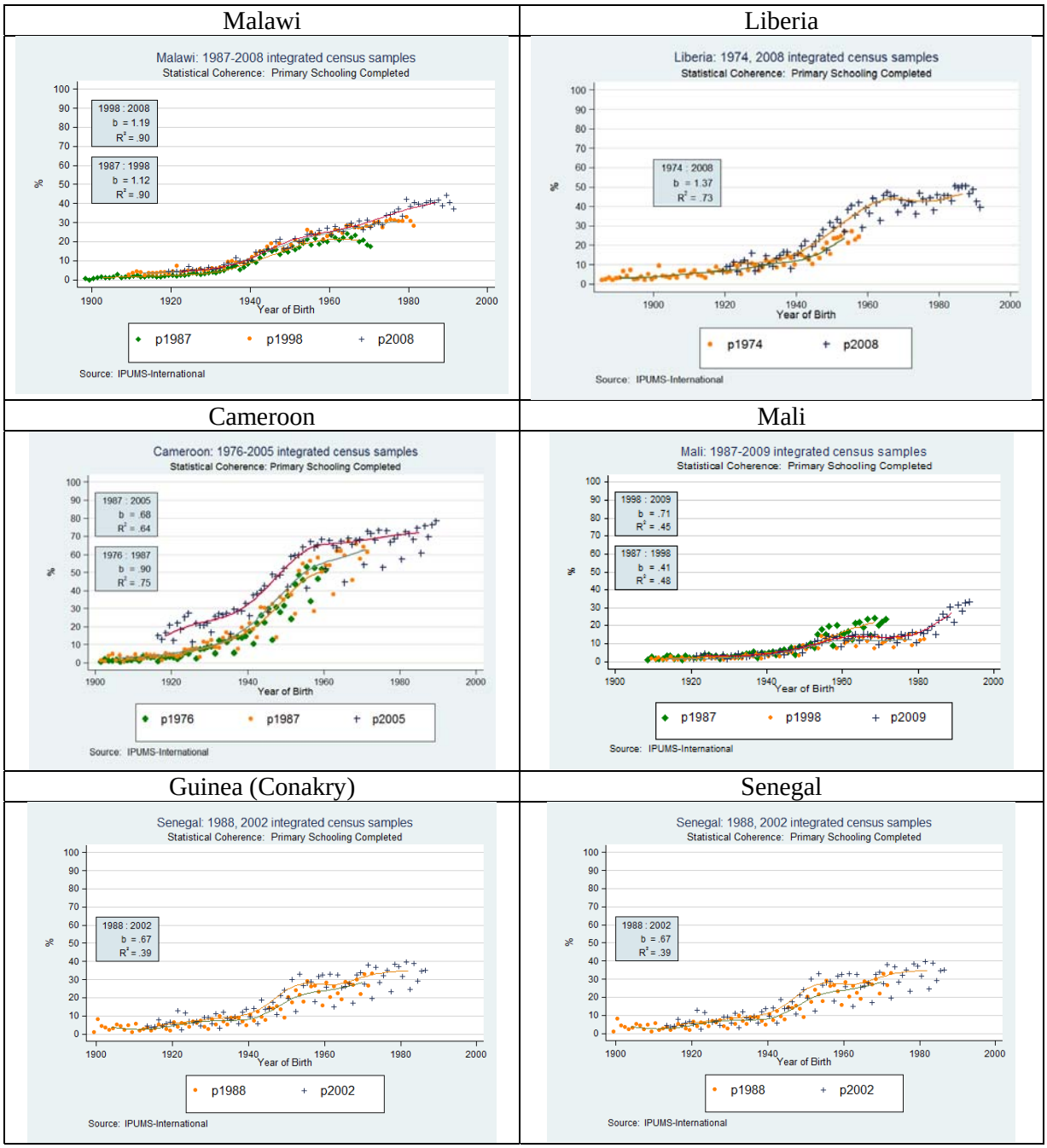


Figure 4. Statistical coherence of primary schooling completed in IPUMS integrated census samples of twelve African countries (continued)
Notes.

[1] https://international.ipums.org/international/variance_estimation.shtml
[2] https://international.ipums.org/international-action/variables/EDATTAN#comparability_section
[3] For a recent comprehensive discussion of technical and political aspects of the 2006 census of Nigeria, see Obono and Omolaubi (2014).
[4] https://bibliography.ipums.org/citations/search with keyword “Mali” and project “IPUMS-International” selected.
[5] https://ec.europa.eu/CensusHub2.