Statistical coherence of primary schooling in IPUMS-International integrated population samples for China, India, Vietnam, and ten other Asia-Pacific countries[1]

Authors:
Robert McCaa, Lara Cleveland, Patricia Kelly-Hall, Steven Ruggles, and Matthew Sobek[2]

Abstract.

IPUMS-International www.ipums.org/international disseminates harmonized census microdata for more than 80 countries at no cost, although access is restricted to bona-fide researchers and students who agree to the stringent conditions of use license. Currently over 270 samples are available, totalling more than 600 million person records. Each year 15-20 additional samples are released, as more countries cooperate with the IPUMS initiative and the integration of 2010 round census samples is completed. With so much microdata so readily available, questions of data quality naturally arise. This paper focusses on the concept of statistical coherence over time for a single concept, primary schooling completed. From an analysis of the percentage completing primary schooling by birth year for pairs of samples for thirteen Asia-Pacific countries, we find outstanding coherence for four—China, Mongolia, Vietnam, and Indonesia—with mean differences of less than 0.5 percentage points, regression coefficient (b) ranging from 0.93 to 1.07 and R^2 =.99. For the thirteen countries as a group there is considerable variation overall with mean absolute difference as high as 16 percentage points, b ranging from 0.62-1.44 and R^2=.65-.99. As a whole,

---

[1]Research for this paper was funded in part by the National Institutes of Health of the United States of America, grant HD047283 European and Asian census microdata harmonization project (IPUMS-EurAsia). The authors express gratitude to the statistical offices that entrusted the original source microdata for integration into the IPUMS-International database and for rights to disseminate extracts to researchers worldwide at no cost without regard to nationality, country of birth or residence. The Asia-Pacific statistical offices are: Bangladesh Bureau of Statistics; National Institute of Statistics, Cambodia; National Bureau of Statistics, China; Bureau of Statistics, Fiji Islands; Ministry of Statistics and Programme Implementation, India; BPS Statistics Indonesia; National Statistical Committee, Kyrgyz Republic; Department of Statistics, Malaysia; National Statistical Office, Mongolia; Statistics Division, Pakistan; National Statistics Office, Philippines; National Statistical Office, Thailand; and General Statistics Office, Vietnam. We thank the reviewers for their many helpful comments and suggestions. The authors alone are solely responsible for errors of analysis or interpretation. A version of this paper was presented at the 27th Population Census Conference (ANCSDAAP), Tokyo, Japan, 5-7 November, 2014.

[2] Minnesota Population Center, Minneapolis, MN USA

Corresponding author:
Robert McCaa, Minnesota Population Center, 50 Willey Hall, 226 19th Ave. S., Minneapolis, MN 55455, USA
Email: rmccaa@umn.edu

statistical coherence of primary schooling is outstanding. Nonetheless, to make expert use of the harmonized microdata, researchers are cautioned to carefully study the IPUMS integrated metadata as well as the original source documentation. National Statistical Offices not currently cooperating or that have not yet entrusted 2010 round census microdata are invited to do so.

Keywords: primary schooling, statistical coherence, IPUMS-International, population census samples, integrated microdata, microdata access, China, India, Vietnam, Asia, Pacific, Bangladesh, Cambodia, Fiji Islands, Indonesia, Kyrgyz Republic, Malaysia, Mongolia, Pakistan, Philippines, Thailand

Introduction

The IPUMS-International database, now in its fifteenth year, disseminates more than 250 integrated census samples representing some 80 countries to researchers across Asia, the Pacific, and the entire globe. The National Bureau of Statistics (NBS) of China is a founding member of the initiative, having endorsed the project memorandum of cooperation in 2002. In 2003, the first Chinese sample, for the 1982 census, was integrated into the database and was followed a few years later by a one percent sample of the 1990 census. An integrated, high precision sample of the 2000 census is planned for launch in 2016. As of this writing, no sample for the 2010 census has yet been made available.

Microdata for all countries in the IPUMS database are disseminated at no cost, but they are not "open data" or "public access." Access is restricted to researchers and policy makers who agree to the stringent conditions-of-use license. Currently, approved users of more than 130 nationalities may access over 615 million person records representing four-fifths of the world's population. By 2020, the database is likely to double with the integration of samples not only for the 2010-round of censuses but also for the backlog of other countries with microdata already entrusted to the Minnesota Population Center. In addition, further expansion is likely as National Statistical Offices (NSO) not yet cooperating with the project decide to do so (Table 1). Appendix 1 provides additional information about the IPUMS-International project.

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** Belarus, **Kenya, Kyrgyz Republic, Mali**	**2009** Guinea Bissau	**2005** Bhutan, Kuwait, Laos, United Arab Emirates; **2006** Hong Kong SAR, Libya, Macau SAR, Maldives, Nigeria (NPC); **2007** Congo Republic, French Polynesia, Swaziland; **2008** Algeria, Burundi, Korea DPR; **2009** Azerbaijan, Chad, Djibouti, Kazakhstan, New Caledonia, Solomon ernatio
2010	**Argentina, Brazil, Dominican Republic**, **Ecuador, Ghana, India (NSSO), Indonesia, Mexico, Panama, Puerto Rico**,Trinidad and Tobago, **USA, 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**, Botswana, Czech Republic, **France,** Greece, Hungary, Iran, **Ireland,** Namibia, **Nigeria (NBS)**, Poland, **Portugal,** Romania, **South Africa, Spain, Uruguay**	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, India (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

With IPUMS managing access to so much microdata for so many countries, the issue of data quality becomes paramount. This paper addresses quality by comparing statistics from selected pairs of samples. We test the statistical coherence of educational attainment, specifically primary schooling from two successive samples for thirteen Asia-Pacific countries (Minnesota Population Center 2014). Because the Chinese census microdata are so important, for China, we extend the analysis backward to a third sample, to 1982, and forward to 2010 in the hope that the NBS will entrust a sample of its latest census in the not-too-distant-future. The analysis of the 2010 census uses a be-spoke table obtained from the NBS website (National Bureau of Statistics 2012).

Conceptual Issues.

For both researchers and NSOs, questions of quality of the samples disseminated by IPUMS-International are of great concern. 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 analyse a single aspect of quality—coherence—for a single dimension—over time[3]—and a single indicator—primary schooling. We ask the question, how do sample statistics for primary schooling completed from the most recent census sample (the predictor variable) compare with an earlier census (the response variable) for the same country? In other words, for each pair of censuses, we backcast the percent completing primary in the most recent sample to a prior sample. The question is not whether the most recent census is the most accurate, but instead to what degree is each pair coherent. We test primary schooling not only because universal primary education is a Millennium Development Goal but also because it is measured by most Asian and Pacific censuses and is widely available in samples disseminated by IPUMS-International.

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 pairs of samples are then compared. Where statistics are coherent from one census to the next, they will show the same or closely similar percentages for primary schooling, birth year-by-birth year. In addition to mean and median absolute differences, we use the correlation coefficient R-squared (R^2) and the least-squares regression coefficient (b) to measure the degree of coherence for each pair of samples and limit the analysis to a constant 55 years of birth for the sake of comparability for all countries.

The results of the analysis are quite promising, indicating a remarkable, even outstanding, degree of coherence with mean and median differences of less than 0.5 percentage points, for 2000 and 2010 census rounds as we will show for China (R^2=.99, b=.95), Vietnam and Indonesia (R^2=.99, b=.93). For earlier rounds, we find similar patterns of coherence for China (comparing 1990 and 1982 with 2010), Mongolia, and Thailand. Nonetheless the range for all thirteen countries is wide (R^2=.65-.99, b=.62-1.44), suggesting significant variations in statistical coherence in successive pairs of samples. Our study of

[3] Four dimensions of coherence—within a dataset, across datasets, over time, and across countries—are identified by the OECD (2011/1: 10).

African census microdata reports somewhat greater dispersion with R^2 ranging from .38 to .99 and b from .46 to 1.37 (McCaa et al. 2015).

Population census data are collected by nations at great expense and have enormous capacity to inform public policy. 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 manner that maximizes their potential.

The Sixteenth Meeting of the United Nations Economic Commission for Europe Group of Experts on Population and Housing Censuses defines coherence as follows (UNECE 2014: 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). To achieve statistical coherence, definitions, concepts, frameworks and classifications must be clear and consistent at the national and international levels. When these change, explanatory text is 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, 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).

Data and methods

A population census is embedded with the demographic history of a nation and its people. Successive, high quality national censuses should tell similar, coherent stories. The population historian's tool kit (four of the authors are population 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 reported completing primary school in the most recently available sample similar to that for one from ten years or so earlier? With Vietnam as an example, we ask: Is the proportion graduating primary or higher schooling of those born in, say, 1970 the same in the sample for the 2009 census as for the census of 1999? As a matter of fact, the answer is yes, almost exactly: 73.8% of Vietnamese born in 1970 completed primary or higher schooling

according to the 2009 census sample compared with 72.5% in the 1999 sample. The difference is scarcely more than a single percentage point.

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. For Vietnam 1999 compared with 2009, as seen in Figure 1 below, we find b=.93, which indicates a high degree of coherence although not a perfect 1.0. R^2 is perfect at .99.

There are at least three caveats for assessing external coherence as proposed hear: 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 data integrated by IPUMS-International, we must also consider how the IPUMS team harmonized the microdata, and whether decisions taken to integrate coding schemes in successive censuses were correct or not. Third, the method assumes that there are no differentials in mortality, migration or reporting by level of educational attainment. The method also assumes that no adult education campaigns were undertaken that might increase the percentage graduating primary school after the normal age. Where the less educated suffer from higher risks of dying a systematic upward bias will emerge. Likewise where the likelihood of migration into or out of a country is associated with educational attainment, then lack of coherence will be exaggerated by international movements unrelated to quality of census operations. Likewise, there may be bias in reporting by the respondents, particularly where educational attainment is low and ages are reported in rounded figures, such as zero, five, etc. For additional details on the method see Feeney (2014).

Integrating educational attainment – the IPUMS-International approach. The principal benefit of IPUMS-International to researchers and NSOs alike is the integration of several decades of microdata samples for each country—typically beginning with the earliest census for which microdata exist or are recoverable and continuing through the 2010 round and beyond. When the project began, few NSOs disseminated census samples. Today most do. Nonetheless, even today, few NSOs publish documentation to facilitate comparative analysis of two or more census samples. Even fewer re-examine earlier censuses to produce cross-walk tables for researchers to aid in harmonizing variables in successive censuses. 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 distribute. Often, little guidance is offered on how to compare microdata from successive censuses.

IPUMS-International systemically collects, archives and disseminates original source documentation, including enumeration forms and instructions to field enumerators as well as codebooks, technical manuals, and official publications. Using the full panoply of such documents, we integrate high-precision microdata, sample by sample, variable-by-variable, and code-by-code. Serial codes in the original microdata are recoded into hierarchical or composite codes to facilitate comparison, yet retain all significant variations in the original

data (Esteve and Sobek, 2003). Integrated metadata are written based on the meticulous study of comprehensive original source documentation. The IPUMS-International team writes metadata for six types of information for each integrated variable in the database (see "tabs" in 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 basic goal of integration 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) have identical codes in every sample. 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 discernible.

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 the samples. Trailing digits provide detail that is only rarely available. A zero place-holder is assigned where information is not available at the level of a particular digit.

For our analysis of coherence, we focus on the IPUMS-International educational attainment variable, "EDATTAN," the single most widely-used variable in the database. Most census microdata with information on this measure follow the UNESCO ISCED scheme (2012), with four levels or stages: whether the respondent completed (a) no schooling at all, (b) primary, (c) secondary or (d) 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 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 illustrates the general and detailed coding schemes for the educational attainment variable for thirteen countries (represented by the two-digit ISO 3166 country code). As the upper section of the table 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 sample-to-sample and country-to-country regarding the various levels of schooling. The frequencies in each cell refer to the simple, un-weighted counts for the corresponding code and sample. The counts are wholly descriptive. As we shall see in the next section, coherence can be assessed by using weighted percentages of the codes, cross-tabulated by year of birth.

Table 2. EDATTAN (educational attainment): IPUMS-International general and detailed harmonized codes for 13 countries
Cell counts refer to un-weighted frequencies for the corresponding codes for the most recent sample integrated

Code	Label	Country (ISO 3166) / Census year	BD 2011	KH 2008	CN 1990	FJ 2007	IN 2004-5	ID 2010	IR 2006	KG 2009	MY 2000	MN 2000	PH 2000	TH 2000	VN 2009
General															
0	NIU (not in universe)		1,117,354	136,274	1,418,185	·	·	2,253,453	131,235	72,044	·	35,396	935,577	43,640	1,517,591
1	Less than primary completed		3,216,705	766,314	4,383,067	24,403	316,386	6,117,917	195,404	124,184	223,334	84,105	2,132,120	284,685	4,675,806
2	Primary completed		2,065,976	376,009	5,069,640	40,755	172,721	10,135,303	467,961	66,697	166,637	54,743	1,967,457	179,347	6,140,145
3	Secondary completed		639,020	47,837	915,562	17,684	85,023	4,394,068	195,055	251,330	10,486	55,050	1,689,518	69,705	1,316,274
4	University completed		166,665	13,010	49,493	1,468	28,290	702,308	59,970	48,606	25,456	14,431	305,054	20,933	527,774
9	Unknown		·	677	·	13	413	·	250,200	2,125	9,387	·	388,084	6,209	·
Detailed															
0	NIU (not in universe)		1,117,354	136,274	1,418,185	·	·	2,253,453	131,235	72,044	·	35,396	935,577	43,640	1,517,591
100	LESS THAN PRIMARY COMPLETED		·	·	·	·	·	·	·	56,168	·	40,263	·	·	·
110	No schooling		1,961,034	297,550	2,145,035	10,890	220,227	1,986,754	41,776	·	100,909	·	466,783	55,479	892,633
120	Some primary		1,255,671	468,764	2,238,032	13,513	96,159	4,131,163	153,628	·	122,425	·	1,665,337	229,206	3,783,173
130	Primary (4 years)		·	·	·	·	·	·	·	68,016	·	43,842	·	·	·
	PRIMARY COMPLETED, LESS THAN SECONDARY														
	Primary completed														
211	Primary (5 years)		1,256,266	·	·	·	88,352	·	233,865	·	·	·	·	·	·
212	Primary (6 years)		·	256,570	2,822,479	24,932	·	6,539,863	·	·	80,005	·	1,967,457	116,450	2,671,203
	Lower secondary completed														
221	General and unspecified		809,710	119,439	2,247,161	15,823	84,369	3,595,440	234,096	46,187	86,632	47,742	·	62,897	3,468,942
222	Technical track		·	·	·	·	·	·	·	20,510	·	7,001	·	·	·
	SECONDARY COMPLETED														
	General or unspecified track														
311	General track completed		639,020	41,385	640,916	10,489	49,669	3,592,138	127,008	208,581	8,878	40,677	814,182	24,371	1,074,774
312	Some college/university		·	·	43,450	380	29,237	·	29,805	15,106	·	·	715,722	17,237	151,141
320	Technical track		·	·	·	·	·	·	·	·	·	14,373	·	·	·
321	Secondary technical degree			2,228	148,554	·	·	400,543	38,242	27,643	·	·	·	13,106	90,359
322	Post-secondary technical			4,224	82,642	6,815	6,117	401,387	·	·	1,608	·	159,614	14,991	·
400	UNIVERSITY COMPLETED		166,665	13,010	49,493	1,468	28,290	702,308	59,970	48,606	25,456	14,431	305,054	20,933	527,774
999	UNKNOWN/MISSING		·	677	·	13	413	·	250,200	2,125	9,387	·	388,084	6,209	·

Source: https://international.ipums.org/international-action/variables/EDATTAN#codes_section
Note: "IN 2004-5" refers to the National Sample Survey Organization Schedule 10 sample; all others refer to national census samples.

The goal of the 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 is open access. Only access to the microdata must be restricted, to respect the conditions of use agreed to by all cooperating NSOs. The integrated microdata are tested and enhanced by extensive analysis. The IPUMS team devotes thousands of hours to analyzing, discussing, debating, testing and re-testing until the microdata integration is validated for dissemination to researchers. The process is repeated annually as additional samples are integrated into the database.

Results

Vietnam. Figure 1 portrays primary schooling rates by year of birth, as computed from the IPUMS integrated samples for the 1989, 1999 and 2009 censuses of Vietnam. The curves reveal astonishing coherence, with regression coefficients of .93 and .92 (1999/2009 and 1989/2009, respectively) strikingly 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 tens of thousands of field workers at three different points in time, separated by intervals of ten years. The data were processed and coded using increasingly sophisticated technologies that nonetheless provide many opportunities for error. Furthermore, the figures are computed from integrated variables constructed by the IPUMS team with no knowledge that the microdata would be examined this way. Nonetheless the statistical coherence of the results in Figure 1 is extraordinary. Researchers should take comfort in the remarkable coherence between successive census samples of Vietnam.

Insert IPUMS Figure 1 near here

Figure 1. Vietnam. Three Census Samples Compared: 2009, 1999 and 1989 Outstanding Statistical Coherence in EDATTAN Primary Schooling Completed

External coherence is all the more noteworthy because, while for all three censuses the General Statistics Office uniformly resorted to face-to-face interviews in the field, the sampling strategy evolved as the number of sample areas increased from a mere 80 in 1989 to 122 in 1999 and to several hundred in 2009. Sample densities also varied greatly, shrinking from 5% in 1989 to 3% in 1999 and then expanding five-fold to 15% in 2009[4].

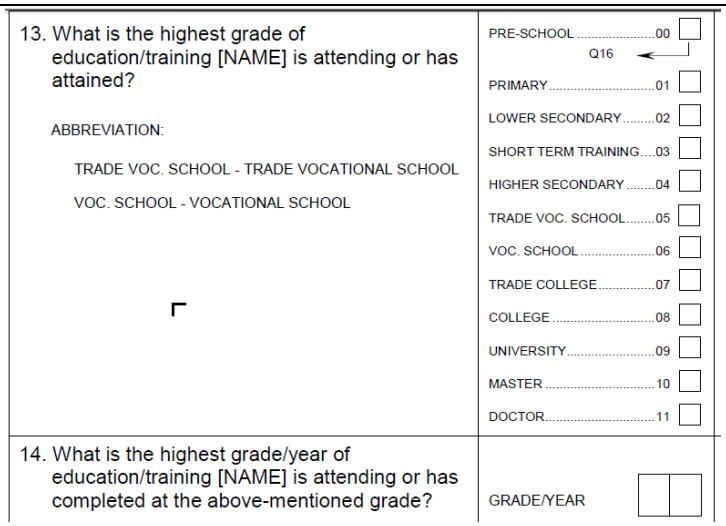
Consider, too, the three caveats referenced above: census agency practices, IPUMS harmonization, and bias.

First, as can be seen in Table 3, the GSO designed questions on educational attainment that, on the whole, are quite consistent from census-to-census. Each census requested both the number of years and level of schooling. Only the 1989 form limited responses to years completed. Regarding attendance, the 1989 and 1999 forms offer three distinct options: whether attending now, attended in the past or never attended. This distinction was dropped from the 2009 form, but this act is of little consequence for our analysis since we are considering only primary education above fourteen years of age.

[4] https://international.ipums.org/international/sample_designs/sample_designs_vn.shtml

Table 3. Educational Attainment Questions in the 1989, 1999 and 2009 Censuses of Vietnam differ in details but are generally quite comparable

2009



13. What is the highest grade of education/training [NAME] is attending or has attained?

ABBREVIATION:

TRADE VOC. SCHOOL - TRADE VOCATIONAL SCHOOL

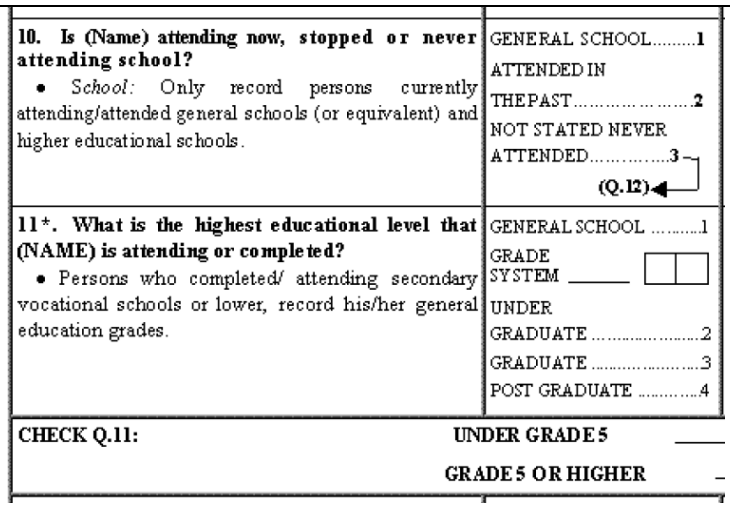
VOC. SCHOOL - VOCATIONAL SCHOOL

PRE-SCHOOL 00 Q16
PRIMARY.....01
LOWER SECONDARY.....02
SHORT TERM TRAINING....03
HIGHER SECONDARY.....04
TRADE VOC. SCHOOL.....05
VOC. SCHOOL 06
TRADE COLLEGE.....07
COLLEGE.....08
UNIVERSITY.....09
MASTER 10
DOCTOR.....11

14. What is the highest grade/year of education/training [NAME] is attending or has completed at the above-mentioned grade?

GRADE/YEAR

1999



10. Is (Name) attending now, stopped or never attending school?

- *School:* Only record persons currently attending/attended general schools (or equivalent) and higher educational schools.

GENERAL SCHOOL.....1
ATTENDED IN THE PAST.....2
NOT STATED NEVER ATTENDED.....3 (Q.12)

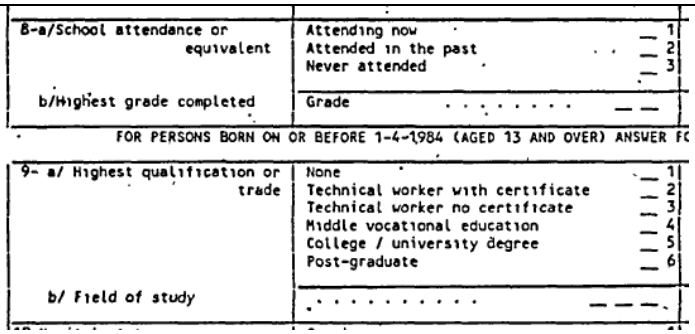
11*. What is the highest educational level that (NAME) is attending or completed?

- Persons who completed/ attending secondary vocational schools or lower, record his/her general education grades.

GENERAL SCHOOL 1
GRADE SYSTEM _____
UNDER GRADUATE 2
GRADUATE 3
POST GRADUATE 4

CHECK Q.11: **UNDER GRADE 5** ____
GRADE 5 OR HIGHER _

1989



8-a/School attendance or equivalent: Attending now — 1; Attended in the past — 2; Never attended — 3

b/Highest grade completed: Grade . . . . . — —

FOR PERSONS BORN ON OR BEFORE 1-4-1984 (AGED 13 AND OVER) ANSWER FC

9- a/ Highest qualification or trade: None — 1; Technical worker with certificate — 2; Technical worker no certificate — 3; Middle vocational education — 4; College / university degree — 5; Post-graduate — 6

b/ Field of study . . . . . — — —

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).

Second, the IPUMS harmonization of educational attainment offers two variants: international and national, EDATTAN, and, in the case of Vietnam, EDUCVN, respectively. For the international recode, only nine codes are needed compared with 95 for EDUCVN. To explain EDUCVN, over 500 words are required for the IPUMS metadata comparability discussion. Yet a mere two sentences suffice to summarize the different thrust of the two variables. The IPUMS metadata reads:

> A harmonized international classification of educational attainment is available in EDATTAN, which imposes a number of compromises to regularize the data across countries. In contrast, EDUCVN retains the full detail on educational attainment from the Vietnam samples.[5]

For completion of primary schooling in the international recode, we have imposed, 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 variables to check consistency and accuracy of the countless decisions made in the harmonization process. The "Source Variable" tab in Figure 2 points the way. For our purposes Figure 2 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 this standard throughout the entire database, even though in the case of Vietnam the National Education System defines primary schooling as completed with only five years of attendance.

Researchers studying only a single country will probably favor (and download) the national recode variable, such as EDUCVN, EDUCBD, EDUCCN, etc., while others interested in comparing differences between countries are likely to pick the international variant, EDATTAN.

Insert IPUMS Figure 2 near here

Figure 2. IPUMS-International Metadata Screen-Grab: Educational Attainment – International Recode Comparability Discussion, Vietnam

In assessing external coherence, the third caveat—bias introduced by assumptions regarding migration, mortality, adult education, or reporting— should also be considered. The fact that the 1999 and 1989 proportions are systematically, if only very slightly lower, at every age may suggest the effects of an adult education campaign, or that the less educated have slightly worse survival chances or higher out-migration rates than the better educated and therefore the proportions with primary schooling completed tend to rise in successive censuses. There may also be an upward bias in reporting events more distant in the past, although in the present case the bias would seem to be very slight.

India. For a second test, we turn to India, where the microdata are not from population censuses but are instead nationally representative samples conducted under the auspices of of the Ministry of Statistics and Planning Implementation by the National Sample Survey Organization ("Socio-Economic Survey, Household Schedule 10"). While Schedule 10 surveys are conducted quinquennially and all are disseminated by IPUMS-International, we examine only three: 1983 (calendar year), 1993 (July, 1993–June, 1994), and 2004 (July, 2004–June 2005).

Insert IPUMS Figure 3 near here

[5] https://international.ipums.org/international-action/variables/EDUCVN#comparability_section

Figure 3. India. Three National Sample Survey Organization Rounds (2004/5, 1993/4, and 1983/4) Show Good Statistical Coherence Despite Severe Age Heaping

Over more than two decades, the NSSO questionnaires maintain a uniform definition of primary schooling completed. However, the forms do not contain a question on years of schooling. Thus, EDATTAN adheres to the Indian national practice: primary equals completion of 5 years; lower secondary, 8 years; and secondary, 10 years. Overall, despite strong digit preference, a high level of coherence is apparent in the summary statistics (R^2=.95, b=1.09, and mean absolute difference = 0.7 percentage points).

Table 4 offers additional statistical detail for assessing coherence for pairs of samples of these and other countries. The law of large numbers might lead one to suspect that larger sample sizes are associated with higher levels of statistical coherence, but this is not the case. China and Pakistan are both characterized by large samples, but only the China samples show high statistical coherence. Mongolia and Thailand are represented by small samples in the IPUMS database but their statistical coherence is outstanding despite their "tiny" size.

Strong digit-preference in age reporting of the uneducated, such as the NSSO samples of India as well as census samples of Bangladesh and Pakistan, distorts any chronological comparison. The Whipple age heaping index reported in Table 4 indicates that age declarations in the NSSO samples are "very rough". Nonetheless we find only a 4.2 percentage point difference for the 1970 birth cohort, with the 2004/5 sample reporting 53.9% completing primary compared with 58.2% for 1993/4. Moreover, the mean difference over the 55-year range common to both surveys is only 0.7 percentage points; the median is even smaller at 0.4.

Table 4. Statistical Coherence in Primary Schooling Between Pairs of Samples For 13 Asia-Pacific Countries

Country	Year of Sample	Sample Size x10^6	Whipple Index & Score	SCHOOL Attendance	YRSCHL Years	EDATTAN Levels	1970 Birth Year %	Mean %	55 Over-lapping Birth Years Difference Mean	Median	R^2	b
				Number of Categories								
Bangladesh	2011- 5%	7.2	262-e	2	14	7	41.4	33.8	-3.1	-2.6	0.93	0.93
	2001-10%	12.4	300-e	3	15	10	42.8	36.9				
Cambodia	2008-10%	0.3	110-b	3	16	10	44.6	28.1	5.4	5.3	0.97	0.93
	1998-10%	0.2	118-b	3	16	8	40.5	22.7				
China	2000- 1%	11.8	100-a	5	-	10	93.9	53.9	0.2	0.0	0.99	0.99
	1990- 1%	11.8	101-a	4	-	7	93.6	54.0	-3.5	-3.4	0.99	0.96
	1982- 1%	10.0	102-a	-	-	8*	91.1	48.7				
Fiji Islands	2007-10%	0.1	105-b	4	15	9	96.6	80.1	6.9	2.9	0.94	1.44
	1996-10%	0.1	102-a	3	15	9	95.6	73.2				
India	2004/5-0.1%	0.6	193-e	5	-	9	53.9	42.7	0.7	0.4	0.95	1.09
	1993/4-0.1%	0.6	221-e	3	-	8	58.2	42.0				
Indonesia	2010-10%	23,6	114-c	4	-	9	86.7	65.2	0.5	0.1	0.99	0.93
	2000-10%	20.1	152-d	-	-	8	84.0	65.8				
Kyrgyz	2009-10%	1.1	100-a	2	-	10	99.1	82.9	1.9	0.9	0.98	0.98
Republic	1999-10%	0.5	99-a	4	-	9	99.1	81.1				
Malaysia	2000- 2%	0.4	115-c	3	-	8	76.1	35.4	-16.1	-13.7	0.93	1.02
	1991- 2%	0.3	114-c	4	15	7	89.2	51.5				
Mongolia	2000-10%	0.2	99-a	3	-	8	94.8	59.1	0.0	0.1	0.99	0.98
	1989-10%	0.2	100-a	-	-	8	92.3	59.1				
Pakistan	1998-10%	13.1	187-e	-	-	9	40.4	22.0	-2.7	-3.6	0.65	0.62
	1981-10%	8.4	264-e	-	-	8	32.6	24.7				
Philippines	2000-10%	7.4	110-b	3	17	9	85.4	65.4	2.3	1.7	0.99	1.04
	1990-10%	6.0	111-b	3	18	9	84.4	63.0				
Thailand	2000- 1%	0.6	110-b	4	24	11	82.4	26.8	0.9	0.7	0.99	1.00
	1990- 1%	0.5	105-b	4	24	12	82.3	25.9				
Vietnam	2009-15%	14.2	101-a	5	23	9	73.8	50.8	0.2	-0.4	0.99	0.93
	1999- 3%	2.4	100-a	5	19	9	72.5	50.6				

Source: www.ipums.org/international *Data do not distinguish between educational level begun or completed.
Note: Whipple index: a. highly accurate (<105), b. fairly accurate (105-109.9), c. approximate (110-124.9), d. rough (125-174.9) and e. very rough (>=175) data. Computations by the authors. IPUMS variable names are indicated by full-caps.

China. **N**ear perfect coherence is apparent for the microdata for the 1982, 1990, and 2000 censuses of China, despite the fact that the 1982 census did not adhere precisely to international standard definitions of educational attainment. The 1982 census did not distinguish between graduates from non-graduates (those who began a particular level of schooling but did not complete it). Therefore the 1982 census microdata do not make the distinction either. Nonetheless educational attainment statistics can be harmonized by re-grouping tabulations from the 1990 and 2000 microdata to match the 1982 definition as we have done in Table 4. The results demonstrate truly outstanding statistical coherence among the three samples.

Comparing 1990 with 2000 both the correlation and regression coefficients are nearly perfect—.99 with a mean difference of 0.2 percentage points. The median difference is exactly zero. These are the best examples of statistical coherence for any of the country comparisons in this paper. The 1982 comparison is outstanding, although not perfect, with a regression coefficient of "only" .96 and a mean difference almost thirty times greater than for the 1990:2000 comparison. However even for 1982 the mean difference is only -3.5 percentage points.

The NBS is to be commended for strictly adhering to international standards for educational attainment definitions for all subsequent population censuses of China. The 2000 and 2010 censuses, with three questions on education (level, status, and continuing adult) are the most detailed and up-to-date with international best practices, while the 1982 census, in comparison, offers scant information with only a single question on educational level and without distinguishing between graduates and non-graduates.

In the absence of microdata for the 2010 census, statistical coherence can be compared with the 1982-2000 microdata by adopting a slightly different concept: instead of "graduated", we adopt "more than graduated". In other words, we shift from "Primary completed" to "additional schooling beyond primary". The concept becomes not simply the completion of primary schooling but actual enrolment in schooling at a higher grade than primary. If the 2010 census microdata were available, it would be possible to harmonize codes at all levels, including primary completed. With only tabular data such coherence is not readily achieved because tables are constructed using fixed categories and definitions. For the 2010 census it is possible to obtain a table of educational attainment from the NBS website[6] that, for some levels of educational attainment, such as beyond primary completed, can be made to harmonize with definitions in the microdata samples from earlier censuses.

Insert IPUMS Figure 4 near here

Figure 4. China. Outstanding Statistical Coherence of Four Censuses: 1982, 1990, 2000, and 2010

For China, as shown in Figure 4, statistics from the 1982-2010 censuses on "more than primary education" are highly coherent overall. Using the 2010 census as the yardstick, the regression coefficients at .95 are exactly identical for 1982, 1990 and 2000—and very close to the .99 as reported in Table 4 and the ideal of 1.0. The product moment correlation coefficients are nearly perfect also at .98 for 1982 and .99 for 1990 and 2000. For the percentage comparison

[6] http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm (accessed March 26, 2015)

we use the birth year 1960 to allow for sufficient years prior to the 1982 census. We find 64.8%, 66.1%, 68.6%, and 69.8% for the 1982-2010 censuses, respectively. The overall mean for the 39-year period, 1925-1963, rises from 30.0% for the 1982 census to 32.2 for 1990, 34.4 for 2000 and 35.7% for 2010. This rise is consistent with higher life expectancies for the more educated, but it is also consistent with additional education at older ages such as adult education courses The fact that the discrepancy is sharpest for younger cohorts than for older ones suggests that the difference is likely attributable to adult education. Adult higher education enrolment grew 3.5% annually from 1978-1998, then jumped to 10.4% per year through 2010 (Lai 2014: 62). We are confident that the results for the 2010 census microdata would be even more coherent than for the table because it would be possible to tabulate the microdata by "rite of passage", such as "primary completed," instead of by the strange, non-intuitive concept of "more than primary".

Additional country comparisons. Table 4 summarizes the analysis for thirteen Asia-Pacific countries with pairs of samples currently disseminated by IPUMS-International. Nearly perfect coherence is attained by three—China, Mongolia, and Thailand. This group shows a mean difference of less than one percentage point, b <+/-0.2 deviation from unity. A second group, with mean differences slightly greater and coefficients slightly lower, characterize pairs for Bangladesh, Cambodia, India, Indonesia, Kyrgyz Republic, Philippines and Viet Nam. A third cluster shows substantial mean differences, from six to sixteen percentage points, yet the coefficients are quite reassuring. Finally, the figures for Pakistan constitute an anomaly due to the extreme age-sex heaping in the overall population structure. On the one hand the mean absolute difference at -2.7 percentage points is relatively small, on the other the regression and correlation coefficients are both below 0.7.

Discussion.

Coherence over time in successive census samples, as measured by the intra-cohort comparison method, is a strong statistical test. Nonetheless coherence over time is rarely assessed because the method is difficult to apply unless the microdata are accessible and integrated. Once integrated into a single database, such as in the case with the samples disseminated by IPUMS-International, the method is easily applied for variables characterized by a "rite-of-passage," such as educational attainment, ever-married, children ever born, etc.

Researchers should understand that the IPUMS-International integrations are performed *ex-post-facto.* The National Statistical Agencies—owners of the microdata—are not responsible for the decisions taken by the IPUMS team to design the integration nor for the harmonized codes. In contrast Eurostat's Census Hub dissemination platform was constructed by European statistical offices before the 2010-round of censuses and was begun so that integration was achieved prior to the actual taking of the censuses.[7]

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

[7] https://ec.europa.eu/CensusHub2

The Minnesota Population Center expresses its gratitude to the NSOs of the Asia and Pacific Region that have endorsed the IPUMS-International Memorandum of Cooperation and have entrusted high-precision census samples to the initiative. Census microdata pose challenges for statistical offices with many priorities and a large public with limited use for such specialized information. Cooperating with IPUMS to disseminate integrated international census microdata offers substantial advantages at minimal cost or risk. Statistical offices are relieved of many of the most burdensome tasks and responsibilities for anonymizing and documenting samples. The isolated statistical office that disseminates census microdata on an *ad hoc* basis incurs substantial risks as well as significant costs in human resources—all for a relatively small return with respect to users. The IPUMS project offers important economies of scale in anonymizing, integrating and managing the dissemination of series of census microdata under uniform protocols and with stringent safe-guards, while maintaining the highest standards of quality and coherence.

The response by researchers to globally integrated microdata is illustrated by Figure 5, a global map of the number of registered users by country (and within the USA by state). More than 10,000 researchers have registered to access IPUMS-International integrated microdata, representing more than 130 nationalities and almost 2,000 institutions, including NSOs, universities, United Nations agencies, and research centers (United Nations Population Division, World Bank, World Health Organization, OECD, National Institute for Policy Studies – Japan, etc.). The IPUMS-International bibliography lists more than a thousand citations (https://bibliography.ipums.org). For Asia, China tops the list with 64 citations. This is remarkable because the most recent integrated microdata are now historical, a full quarter century old. India ranks second with 58, followed by Vietnam (49), Philippines (36), and four countries with around two dozen citations each: Cambodia, Indonesia, Malaysia, and Thailand.

Insert IPUMS Figure 5 near here

Figure 5. IPUMS-International Registered Researchers by Nationality and Status of Cooperation (January 1, 2015)

For the 2020-round of population censuses, statistical coherence is likely to be even greater than for the 2010-round thanks to the ever-increasing cooperation between the United Nations Statistics Division, and—most importantly—official statisticians of the National Statistics Offices.

References.

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.

Lai, Qing. 2014. "Chinese Adulthood Higher Education: Life-Course Dynamics under State Socialism." *Chinese Sociological Review* 46:55-79.

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., L. Cleveland, P. Kelly-Hall, S. Ruggles and M. Sobek. 2015. "Statistical coherence of primary schooling in population census microdata: IPUMS-International integrated samples compared for fifteen African countries." *African Population Studies* 29(1):157-180.

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.

National Bureau of Statistics. 2012. *2010 Population Census of China: Tables*, educational level by age and sex. http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm (accessed March 26, 2015).

Organization of Economic Cooperation and Development (OECD). 2011. *Quality Framework and Guidelines for OECD Statistical Activities*. Version 2011/1

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.

Sobek, M and S. Kennedy. 2009 "The development of family interrelationship variables for international census data." Minnesota Population Center. https://international.ipums.org/international/resources/misc_docs/pointer_working_paper_2009.pdf .

United Nations European Economic Commission (UNECE). 2014. Group of Experts on Population and Housing Censuses. "Quality Management" – Draft Text for the *Conference of European Statisticians Recommendations for the 2020 Census Round*. Geneva, Sep. 23-26.

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.

Appendix 1. IPUMS-International Value-Addeds www.ipums.org/international

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. IPUMS-International disseminates microdata on a restricted access basis to a tiny, but important constituency—researchers, such as readers of this journal, who require detailed data on individuals and households to measure and analyse complex relationships, often making comparisons over time and between nations.

IPUMS never disseminates the original, raw source files. Instead the microdata are transformed, harmonized, and integrated such that any single concept, such as primary schooling completed, has the same code in every sample through-out the entire database (see section above, "Integrating Educational Attainment"). Nor are entire datasets disseminated. Instead each researcher constructs by means of electronic menus custom extracts, tailored as to country(ies), census year(s), subpopulation(s), and variables, according to the individual needs of the researcher. Each extract is a single pooled dataset that is registered to facilitate replicability and to guard against fraud. This method provides strong incentives for users to jealously guard the microdata and comply with the conditions of use. Since complete datasets are not distributed on DVDs or any other media, the temptation to share microdata with unauthorized individuals is greatly reduced.

The IPUMS team, with decades of experience in using microdata, developed more than thirty value-added variables that augment each sample. These augmented variables may be grouped into three types: technical, summary and pointer.

- **Technical variables**: Record type, Country, Year, IPUMS sample identifier, Household serial number, Number of person records in household, Household weight, Subsample number, Group quarters status, Continent, Region of country, Residence at first administrative level, and Expansion factors (sample weights—for households and persons).
- **Summary household and family variables**: Household classification, Number of families in household, Number of married couples in household, Number of mothers in household, Number of fathers in household, Head's location in household, Number of unrelated persons, Family unit membership, Number of own family members in household, Number of own children in household, Number of own children under age 5 in household, Age of eldest own child in household, and Age of youngest own child in household.
- **Pointer variables** to identify co-resident spouses, children and their parents: Mother's, Father's and Spouse's location in household, Rule for linking parent(s) and spouse(s), Probable stepmother, Probable stepfather, Man with 2 or more wives linked and Second or higher order wife, etc. This is the one of the most valuable addition to IPUMS integrated samples because it readily facilitates analysis of children by the characteristics of their mothers or fathers as well as husbands by the characteristics of their wives and vice-versa (Sobek and Kennedy 2009). Own-child fertility analysis is made easy because every IPUMS household sample already links mothers to their co-resident children and the "Attach Characteristics" feature of the IPUMS extract system can be used to place mother's characteristics on the record of each child.

Fig 1:

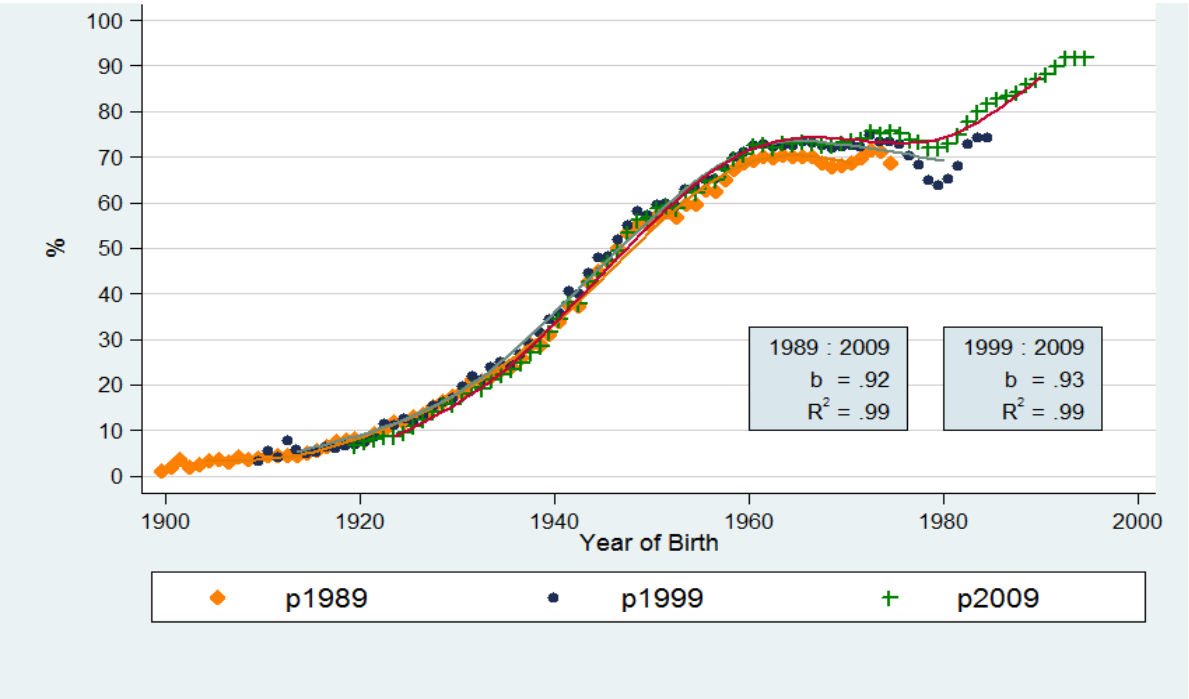
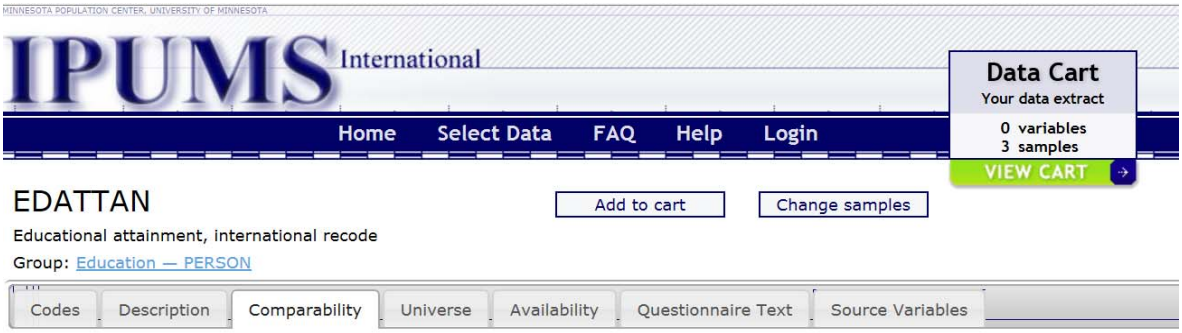

100
90
80
70
60
50
40
30
20
10
0
%
1900
1920
1940
1960
1980
2000
Year of Birth
1989 : 2009
b = .92
R² = .99
1999 : 2009
b = .93
R² = .99
p1989
p1999
p2009

Fig 2.



Comparability — Index

GENERAL Vietnam

Comparability — Vietnam [top]

Vietnam reported individual years of schooling, and therefore fit into the 6-3-3 system used in EDATTAN.

The educational system changed in 2009 to a 5-year primary completion. However, grade-level reporting was used to impose the international 6-3-3 standard for 2009, making educational attainment consistent across Vietnam samples.

In Vietnam 2009, persons with short-term technical training are coded as completing "primary (5 years)"; those with 3 or more years of upper secondary vocational education are considered to have completed secondary; and those with 3 or more years of college or 4 or more years of university are considered to have "university (complete)".

section

Fig 3.

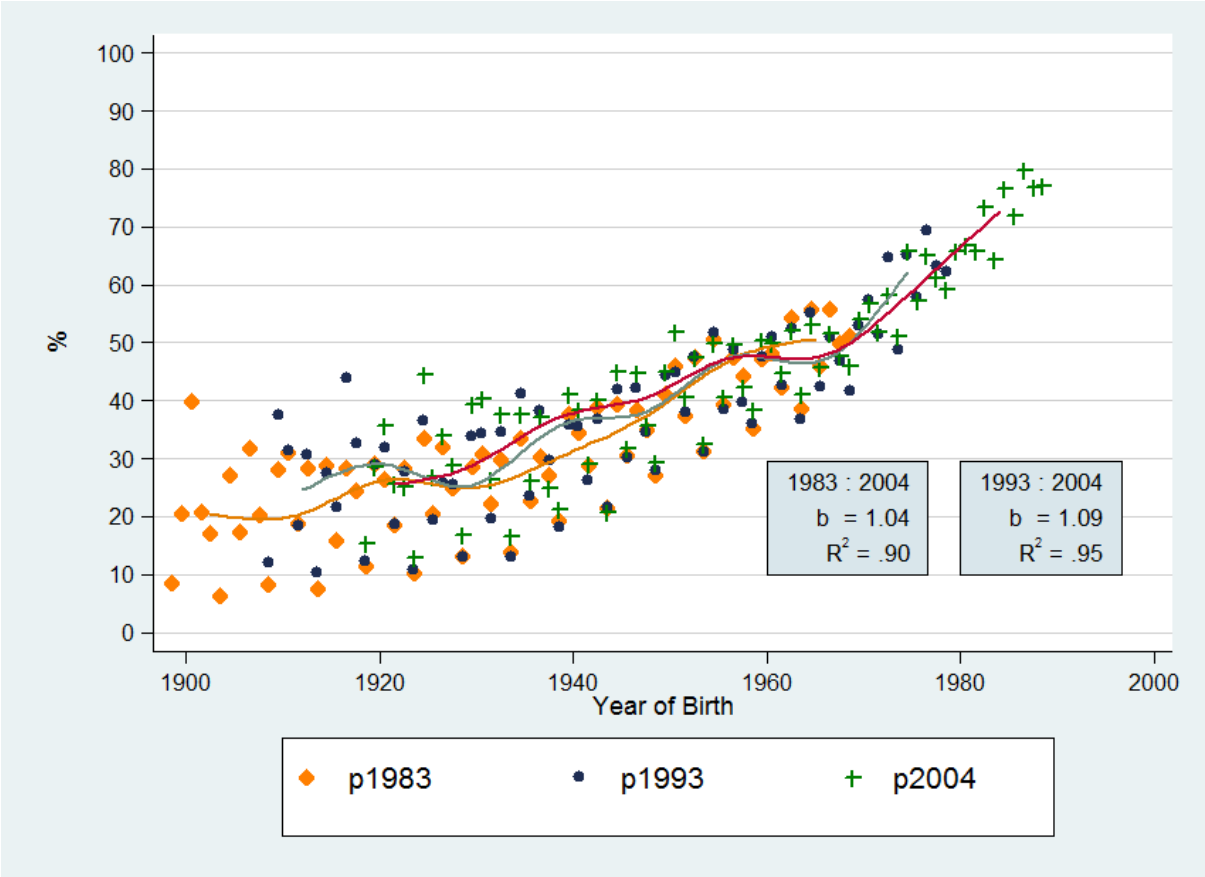

100
90
80
70
60
50
40
30
20
10
0
%
1900
1920
1940
1960
1980
2000
Year of Birth
1983 : 2004
b = 1.04
R² = .90
1993 : 2004
b = 1.09
R² = .95
p1983
p1993
p2004

Fig 4.

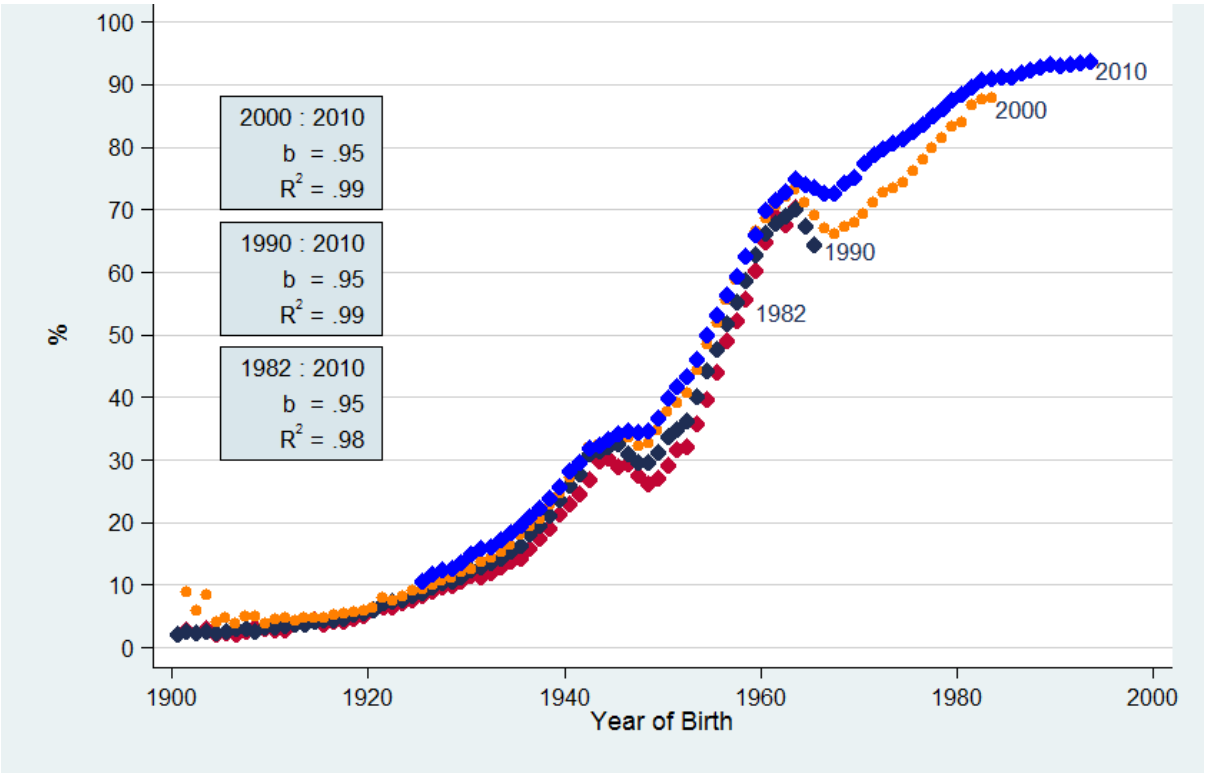
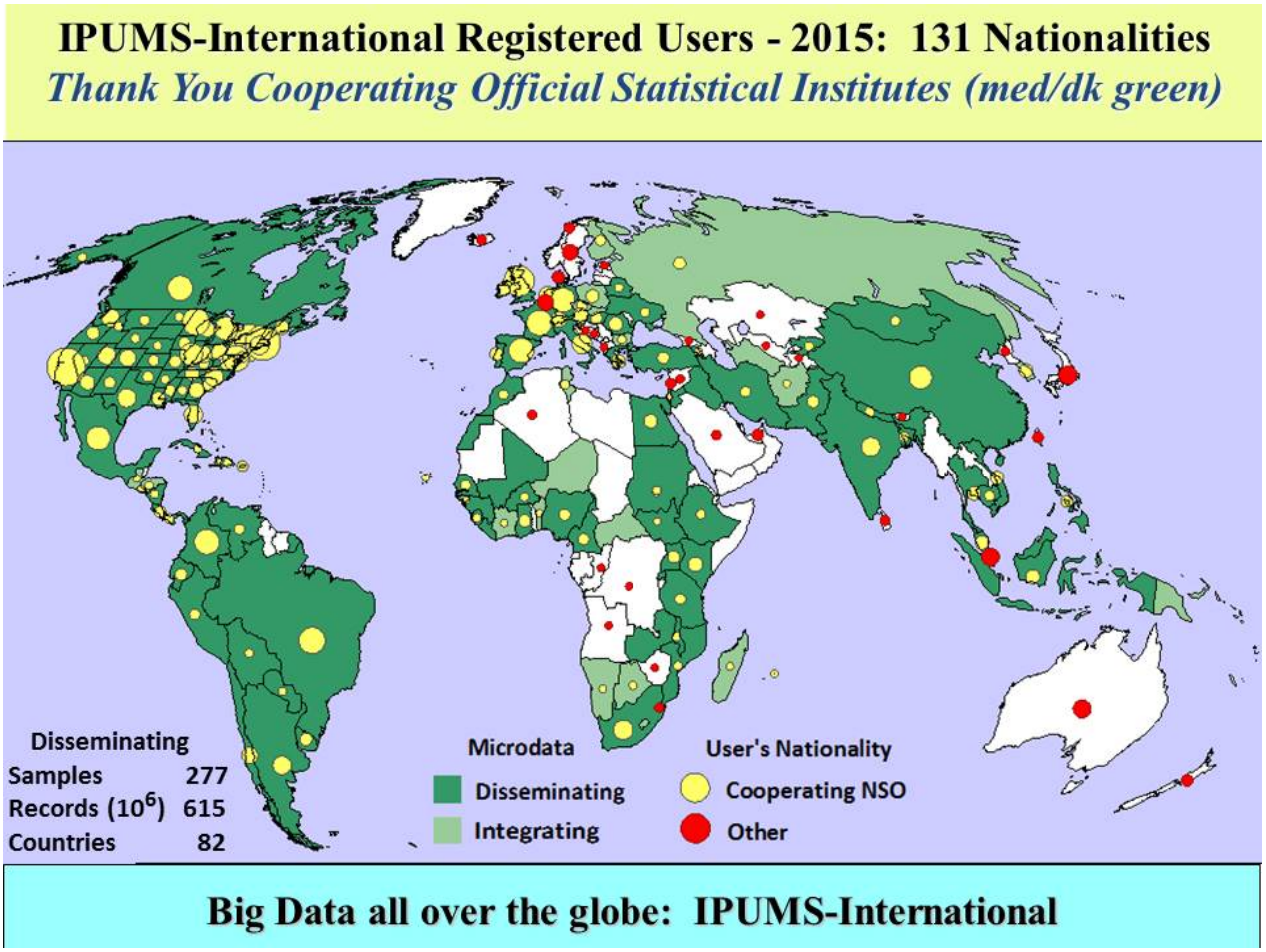

100
90
80
70
60
50
40
30
20
10
0
%
2000 : 2010
b = .95
R² = .99
1990 : 2010
b = .95
R² = .99
1982 : 2010
b = .95
R² = .98
2010
2000
1990
1982
1900
1920
1940
1960
1980
2000
Year of Birth

Fig 5.


IPUMS-International Registered Users - 2015: 131 Nationalities
Thank You Cooperating Official Statistical Institutes (med/dk green)
Disseminating
Samples 277
Records (10^6) 615
Countries 82
Microdata
Disseminating
Integrating
User's Nationality
Cooperating NSO
Other
Big Data all over the globe: IPUMS-International