Comparability of the Public Use Files of the U.S. Census of Population, 1880–1980

STEVEN RUGGLES

MOST OF the information released by the Census Bureau has always consisted of summary population counts cross-tabulated by individual or family characteristics. Although these data form the basic description of the American population, they are not ideal for analytical research. The Census Bureau cannot anticipate all the needs of social scientists, so many topics are inadequately covered in the published census volumes. For historical research, the problem is especially acute, because the published data are fairly sketchy for the period before 1940. Moreover, the classifications employed by the Census Bureau have changed over time, making long-term comparisons difficult or impossible.

In 1963 the Bureau of the Census created the first public use sample (PUS) of the U.S. Census of Population and Housing (U.S. Bureau of the Census 1973). This data file, consisting of separate records detailing the characteristics of 180,000 individuals, was distributed to researchers on punch cards or magnetic tape. To preserve confidentiality, the bureau stripped names and other identifying characteristics from the file. For the first time, researchers were able to make tabulations tailored to their specific

Steven Ruggles is associate professor of history and director of the Social History Research Laboratory at the University of Minnesota. His work with the public use census files has been supported by NIH grant HD 25839, a McKnight Land Grant Professorship, and grants-in-aid from the Graduate School of the University of Minnesota.
Social Science History 15:1 (Spring 1991). CCC 0145-5532/91/$1.50.

research needs. In addition, the 1960 public use sample allowed researchers to move beyond simple tabular analysis and apply increasingly sophisticated multivariate techniques. These microdata proved to be an indispensable resource and immediately led to an outpouring of new research.

For the 1970 census, the Census Bureau released a set of six public use samples, which varied in subject content and geographic detail (U.S. Bureau of the Census 1972a). The density of these samples was increased from the 1 in 1,000 of the 1960 census to 1 in 100, greatly enhancing the potential for study of small population subgroups. In conjunction with the 1970 PUS, the bureau released a new version of the 1960 PUS, enlarged to the 1-in-100 sample density and arranged to simplify comparison with the 1970 census files. The range of subject matter and sample sizes was further increased for the 1980 census. Three public use microdata samples (PUMS) were released; in combination, they provide data on 7% of the U.S. population (U.S. Bureau of the Census 1982a).

In recognition of the value of the series of census microdata files, historical public use samples have been created for earlier census years. Fortunately, the original enumerators' manuscripts survive for all U.S. census years except 1890 (that manuscript burned). Therefore, creation of a new PUS is mainly a task of converting a sample of those manuscripts to machine-readable form. One percent samples of the 1940 and 1950 censuses were constructed by the Census Bureau and the Center for Demography and Ecology at the University of Wisconsin (U.S. Bureau of the Census 1984a, 1984b). Smaller samples from the 1900 and 1910 samples were created under the direction of Sam Preston at the University of Washington and the University of Pennsylvania (Graham 1979; Strong et al. 1989). Finally, a 1-in-100 sample of the 1880 census is now underway here at the Minnesota Social History Research Laboratory (Ruggles and Menard forthcoming). Thus, except for the gaps of 1920, 1930, and 1890, we will soon have a series of microdata census samples covering the past 100 years. Used in combination, the eight datasets spanning a century of cataclysmic social and economic change will constitute our most important resource for the study of changing social structure.

The potential for consistent comparisons across census years

is greatly enhanced by the availability of microdata. Nevertheless, there are significant problems of compatibility across the eight public use samples. The range of questions asked by the census has changed over time, and even where the questions are similar there have usually been changes in census definitions and enumerator instructions. In some areas, social change has been so great that the meaning of certain inquiries and responses has altered. Moreover, the administrative structure of the census and the procedures for gathering the data have evolved, affecting both the completeness of enumeration and the detail of responses. Additional incompatibilities have been introduced in the construction of the public use samples because of inconsistent coding schemes, variations in sampling strategies, and irregular treatment of missing data.

Despite all these problems, the U.S. public use samples constitute the most consistent and comprehensive source there is for the study of long-term social change. As long as we exercise caution in using the samples and carefully consider the potential effects of differences in their construction, the public use samples promise to increase dramatically the power of research on historical social change.

This article is an effort to outline the most important differences among the samples and to suggest strategies for coping with some of the problems of compatibility. A major task of the Social History Data Archives of the University of Minnesota during the past three years has been the development of consistent versions of the public use sample files for 1900 through 1980 for use by graduate students and faculty. As part of this effort, a considerable body of experience with the data has been built up. Space does not permit a full discussion of the compatibility of each variable across all census years, but the importance of the topic demands that we touch on the most problematic issues.

CENSUS FORMAT AND SAMPLE FORMAT

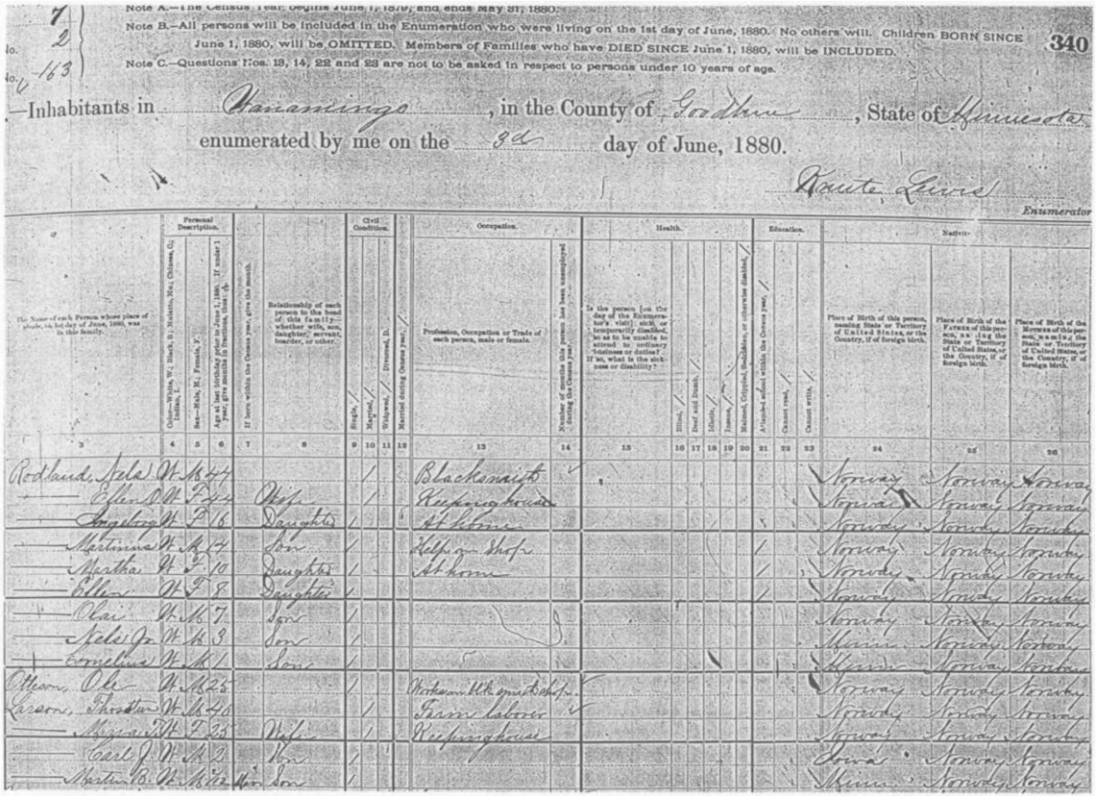
The public use samples are transcriptions of information from the original enumerators' manuscripts. The layout of these census forms has changed in ways that affect the structure and content of the public use samples. Before embarking on a detailed discussion of census definitions, sample designs, and specific variables, I

will therefore briefly describe the major changes in the enumeration schedules and how these changes affected the format of the public use samples.

Prior to 1850, the census office gathered information on households rather than on individuals. Thus, for example, the enumerator asked how many adult women were present in the household instead of asking the age and sex of each individual. This format severely limits the available information and precludes the construction of effective public use samples for the first half century of American history. From 1850 on, the census asked questions about the characteristics of every individual in the population. A reproduction of the census form used in the census of 1880 is shown in Figure 1. The body of the census form is divided into 26 columns, 1 for each question asked. There are 50 lines on each page, and each line contains information on a different individual. The individuals were divided into residential units by giving each dwelling and family a different number in the two leftmost columns of the schedule. Although the specific questions varied, the same basic layout of the enumeration form was used for every census from 1850 to 1930 (U.S. Bureau of the Census 1979).

Because information about both individuals and families is available, the public use samples have adopted a hierarchical structure. Thus, they are simultaneously samples of groups and of the individuals who live in those groups. The names and definitions of the groups vary somewhat among the samples. The public use samples for the censuses of 1940 through 1980 are samples of households, those for 1900 and 1910 are samples of "families," and the 1880 PUS is a sample of dwellings. The next section discusses the implications of these changes in the units of enumeration and sampling.

In all the samples, variables common to the group as a whole, such as geographic information and household structure, are located on a "household record." Each household record is followed by a series of person records giving the characteristics of each member of the household. The analytical power of the public use samples derives largely from this hierarchical organization: within the group, the relationships among individuals are known, and this allows the creation of a wide range of new variables on family relationships, the household economy, generational change, marital unions, and the like.



Note A.—The Census Year begins June 1, 1879, and ends May 31, 1880.
Note B.—All persons will be included in the Enumeration who were living on the 1st day of June, 1880. No others will. Children BORN SINCE June 1, 1880, will be OMITTED. Members of Families who have DIED SINCE June 1, 1880, will be INCLUDED.
Note C.—Questions Nos. 13, 14, 22 and 23 are not to be asked in respect to persons under 10 years of age.

340

Inhabitants in Varianingo, in the County of Goodhue, State of Minnesota,
enumerated by me on the 3d day of June, 1880.

Route Lines
Enumerator

The Name of each Person whose place of abode, on the 1st day of June, 1880, was in this family.	Personal Description. Color—White, W.; Black, B.; Mulatto, Mu.; Chinese, Ch.; Indian, I.	Sex—Male, M.; Female, F.	Age at last birthday prior to June 1, 1880. If under 1 year, state number of months.	If born within the Census year, give the month.	Relationship of each person to the head of this family—whether wife, son, daughter, servant, boarder, or other.	Civil Condition. Single, S.; Married, M.; Widowed, Wd.; Divorced, D.	Married during Census year.	Occupation, Profession, Occupation or Trade of each person, male or female.	Number of months this person has been unemployed during the Census year.	Health. Is the person [on the day of the Enumerator's visit] sick or temporarily disabled, so as to be unable to attend to ordinary business or duties? If so, what is the sickness or disability?	Blind.	Deaf and Dumb.	Idiotic.	Insane.	Maimed, Crippled, Bedridden, or otherwise disabled.	Education. Attended school within the Census year.	Cannot read.	Cannot write.	Nativity. Place of Birth of this person, naming State or Territory of United States, or the Country, if of foreign birth.	Place of Birth of the Father of this person, naming the State or Territory of United States, or the Country, if of foreign birth.	Place of Birth of the Mother of this person, naming the State or Territory of United States, or the Country, if of foreign birth.
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Rodland, Nels	W	M	44					M				Blacksmith											Norway	Norway	Norway
Ellen	W	F	42		Wife			M				Keeping house											Norway	Norway	Norway
Angelina	W	F	16		Daughter			S				At home											Norway	Norway	Norway
Martinus	W	M	14		Son			S				Helps on farm											Norway	Norway	Norway
Martha	W	F	10		Daughter			S				At home											Norway	Norway	Norway
Ellen	W	F	8		Daughter			S															Norway	Norway	Norway
Clara	W	M	7		Son			S															Norway	Norway	Norway
Nels Jr.	W	M	3		Son			S															Minn.	Norway	Norway
Cornelius	W	M	1		Son			S															Minn.	Norway	Norway
Olleson, Ole	W	M	25					S				Works on the mud dock											Norway	Norway	Norway
Larson, Thorsten	W	M	40					M				Farm laborer											Norway	Norway	Norway
Mina	W	F	35		Wife			M				Keeping house											Norway	Norway	Norway
Carl J.	W	M	12		Son			S															Iowa	Norway	Norway
Martin B.	W	M	10	4m	Son			S															Minn.	Norway	Norway

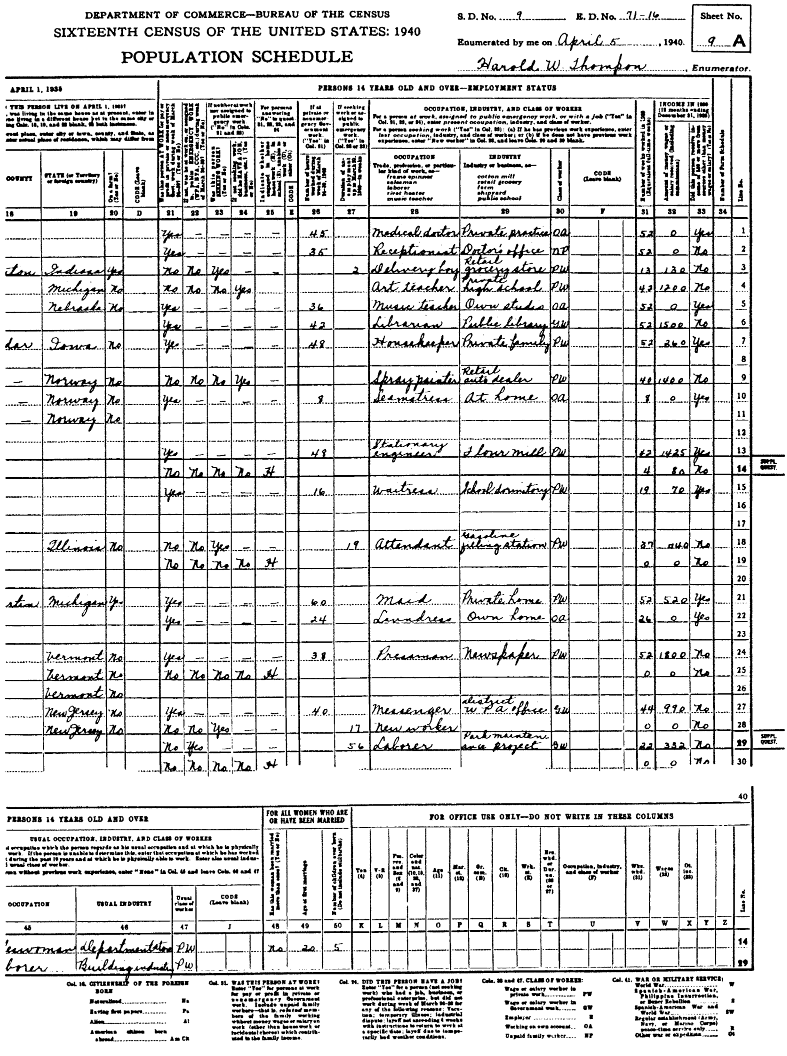
Figure 1 Census form used in 1880

In 1940, the Census Bureau broadened the scope of the census inquiries by asking a set of supplemental questions for a sample of the population. The layout of the form is similar to that of 1850–1930, but two of the rows on each census page are highlighted, and the individuals on those lines were designated "sample-line" individuals. At the bottom of the form there are additional questions to be answered by the persons who happened to fall on the sample lines (Figure 2). The census of 1950 had a similar structure, but the number of questions asked of the entire population was reduced, and more questions were relegated to the sample line.

The public use samples for 1940 and 1950 were designed so that each enumeration unit (household or group quarters) contained one sample-line person. Each household record is followed by a sample-line record giving additional information on the sample-line person. The sample-line record is followed in turn by individual person-records for each member of the household. These data files are greatly enhanced by the availability of the sample questions, but since only one person in each enumeration unit was asked the additional questions, there are limits on the kinds of new variables that can be constructed. For example, the questions relating to ethnic background appear only on the sample line and may be available for either the husband or wife, but never both; thus, one cannot create variables to assess the extent of ethnic endogamy.

The census form was dramatically altered for the census of 1960. The basic structure of the forms from 1850 to 1950—with each column representing a different question and each row representing a different individual—was finally abandoned. Instead, each household received an individual census form, which looked like a multiple-choice examination (see Figure 3). The sample-line questions were eliminated; instead, 25% of households received "long forms," which contained a wide range of additional sample questions. Since the public use sample was constructed entirely from long forms, the additional questions are available for every individual in the file.

The multiple-choice format of the 1960 census was adopted for two reasons. First, the census was largely self-enumerated. Most households received a census form in the mail to be filled out for later collection by an enumerator. The multiple-choice format was



DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS

SIXTEENTH CENSUS OF THE UNITED STATES: 1940

POPULATION SCHEDULE

S. D. No. 9 E. D. No. 71-16 Sheet No. 9 A

Enumerated by me on April 5, 1940. Harold W. Thompson, Enumerator.

APRIL 1, 1935 — PERSONS 14 YEARS OLD AND OVER—EMPLOYMENT STATUS

County (18)	State (or Territory or foreign country) (19)	On a farm? (20)	Code (D)	21	22	23	24	25	Code (E)	Hours worked (26)	Duration of unemployment (27)	Occupation (28)	Industry (29)	Class of worker (30)	Code (F)	Weeks worked in 1939 (31)	Amount of money wages or salary (32)	Other income (33)	Farm Schedule No. (34)	Line No.
				Yes	–	–	–	–		45		Medical doctor	Private practice	OA		52	0	Yes		1
				Yes	–	–	–	–		35		Receptionist	Doctor's office	NP		52	0	No		2
[illegible]ton	Indiana	Yes		No	No	Yes	–	–			2	Delivery boy	Retail grocery store	PW		13	130	No		3
	Michigan	No		No	No	No	Yes					Art teacher	Private high school	PW		42	1200	No		4
	Nebraska	No		Yes	–	–	–	–		36		Music teacher	Own studio	OA		52	0	Yes		5
				Yes	–	–	–	–		42		Librarian	Public library	GW		52	1500	No		6
[illegible]dar	Iowa	No		Yes	–	–	–	–		48		Housekeeper	Private family	PW		52	260	Yes		7
																				8
–	Norway	No		No	No	No	Yes	–				Spray painter	Retail auto dealer	PW		40	1400	No		9
–	Norway	No		Yes	–	–	–	–		8		Seamstress	At home	OA		8	0	Yes		10
–	Norway	No																		11
																				12
				Yes	–	–	–	–		48		Stationary engineer	Flour mill	PW		52	1425	Yes		13
				No	No	No	No	H								4	80	No		14
				Yes	–	–	–	–		16		Waitress	School dormitory	PW		19	70	Yes		15
																				16
																				17
	Illinois	No		No	No	Yes	–	–			19	Attendant	Gasoline filling station	PW		37	840	No		18
				No	No	No	No	H								0	0	No		19
																				20
[illegible]stin	Michigan	Yes		Yes	–	–	–	–		60		Maid	Private home	PW		52	520	Yes		21
				Yes	–	–	–	–		24		Laundress	Own home	OA		26	0	Yes		22
																				23
	Vermont	No		Yes	–	–	–	–		38		Pressman	Newspaper	PW		52	1800	No		24
	Vermont	No		No	No	No	No	H								0	0	No		25
	Vermont	No																		26
	New Jersey	No		Yes	–	–	–	–		40		Messenger	District W P A office	GW		44	990	No		27
	New Jersey	No		No	No	Yes	–	–			17	New worker				0	0	No		28
				No	Yes	–	–	–			56	Laborer	Park maintenance project	GW		22	332	No		29
				No	No	No	No	H								0	0	No		30

PERSONS 14 YEARS OLD AND OVER — USUAL OCCUPATION, INDUSTRY, AND CLASS OF WORKER; FOR ALL WOMEN WHO ARE OR HAVE BEEN MARRIED; FOR OFFICE USE ONLY—DO NOT WRITE IN THESE COLUMNS

Occupation (45)	Usual industry (46)	Usual class of worker (47)	Code (J)	Has this woman been married more than once? (48)	Age at first marriage (49)	Number of children ever born (50)	Line No.
[illegible]eswoman	Department store	PW		No	20	5	14
[illegible]borer	Building industry	PW					29

Figure 2 Census form used in 1940

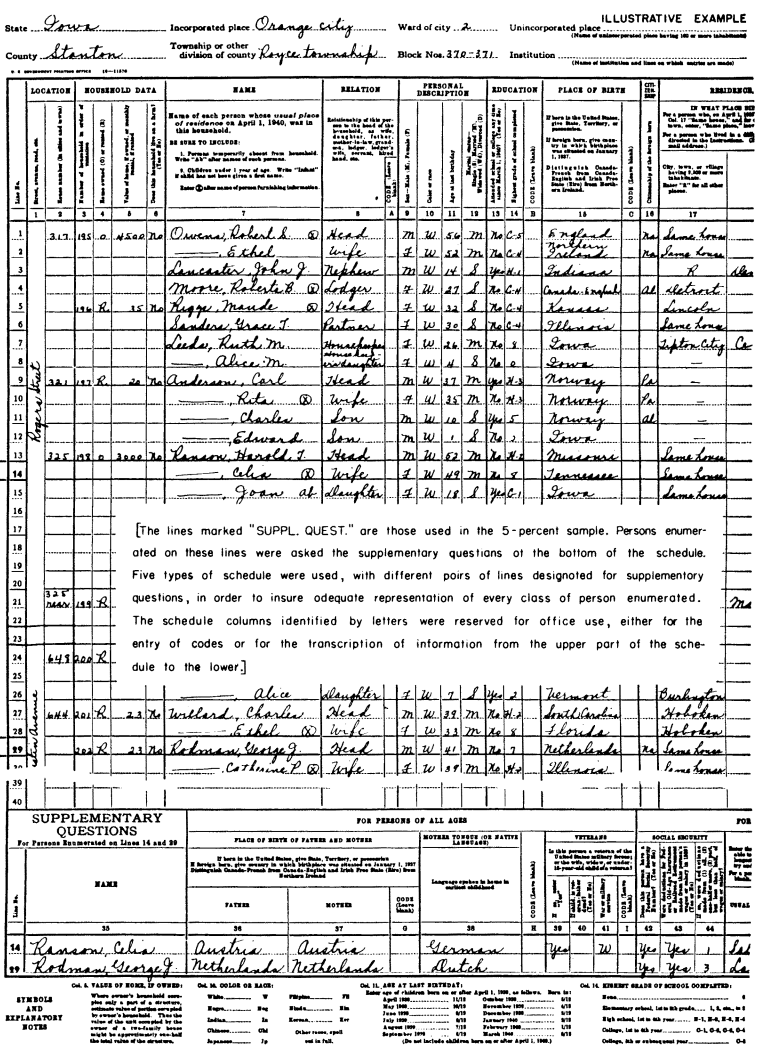


Figure 3 Census form used in 1960

supposed to simplify the task of filling out the form. In addition, the 1960 census was converted to machine-readable form by the Film Optical Sensing Device for Input to Computers (FOSDIC), a machine that can only read little circles filled in with No. 2 pencils (Eckler 1972; Anderson 1988).

The shift to self-enumerated, multiple-choice census forms probably improved the accuracy of responses to some census inquiries. In the public use samples, however, the use of multiple-choice questions also reduced the detail available for several variables. The implications of the change for specific questions are discussed below.

The census forms for 1970 and 1980 were similar to those for 1960. In 1970, the Census Bureau used two long forms with somewhat different questions; one was answered by 5% of households, and the other by 15%. Separate public use samples were constructed from each set of long forms. For the 1980 census, the bureau incorporated all questions into a single long form. In both 1970 and 1980, the Census Bureau released three versions of each public use sample containing alternate geographic codes.

The general characteristics of the public use files are summarized in Table 1. The sample sizes increase steadily with time, with the exception that the 1880 sample will be somewhat larger than the 1900 and 1910 samples. The smaller size of the samples from the early census years limits their usefulness for detailed study of small population subgroups and narrow geographic areas. The 1900 sample, for example, includes only 148 Chinese, 238 persons in Minneapolis, and 266 iron and steel workers.

The three columns on the right of Table 1 give the number of variables on the household record, person record, and sample line of each dataset. These figures provide only a rough guide to the number of census inquiries in each census year, because the public use files vary widely in the number of constructed variables they provide and in the detail of geographic identifiers. As we shall see, for example, the 1880 census provided significantly less information than the 1900 and 1910 censuses, even though the number of variables in the public use sample will be larger for 1880.

Table 1 Characteristics of the public use census files, 1880–1980

Name of file	Density	Approximate number of cases	Record length	Number of variables[a]		
				House-hold	Person	Sample line
1880 PUS	1/100	502,000	120	45	51	
1900 PUS	1/760	100,000	70	33	28	
1910 PUMS	1/250	366,000	111	30	40	
1940 PUMS	1/100	1,317,000	138	30	55	27
1950 PUMS	1/100	1,507,000	133	23	39	42
1960 PUS	1/100	1,793,000	120	63	53	
1970 PUS						
5% state	1/100	2,032,000	120	75	59	
5% county	1/100	2,032,000	120	72	59	
5% neighborhood	1/100	2,032,000	120	73	59	
15% state	1/100	2,032,000	120	63	62	
15% county	1/100	2,032,000	120	60	62	
15% neighborhood	1/100	2,032,000	120	61	62	
1980 PUMS						
A sample	1/20	11,327,000	193	67	78	
B sample	1/100	2,265,000	193	67	78	
C sample	1/100	2,265,000	193	66	78	

[a]Excluding data-quality flags.

CHANGES IN THE UNITS OF ENUMERATION AND SAMPLING

During the century between 1880 and 1980, the basic units of enumeration employed by the census were modified repeatedly. In the censuses of 1880 through 1910, all individuals were assigned to a family. A family was an individual or group of individuals who "jointly occupied" a dwelling place or part of a dwelling place. Census instructions defined dwelling places by the existence of a front door; they included both wigwams and tenement houses. In 1880 and 1900, the number of separate families within a dwelling place was generally determined by the number of separate eating tables. At the discretion of the enumerator, the separate tables requirement could be suspended; the instructions vaguely state that separate meals are "not always" necessary. In 1910, families were distinguished from one another if they occupied separate portions of a dwelling. In all three census years, there

were several additional exceptions to the rules. All the permanent occupants of hotels, institutions, and military barracks constituted single families, provided they slept in the same building. Census enumerators likewise counted boarders, lodgers, and servants as part of the family occupying the dwelling place where they slept, regardless of their eating arrangements. Nonpermanent residents, including hotel guests and students at schools and colleges, were enumerated at their "usual place of abode," which meant that college students in dormitories were supposed to be listed as members of their parental family (the enumeration procedures for this period are documented in U.S. Census Office 1882, 1883, 1895; Walker 1888; Wright and Hunt 1900; Department of Commerce and Labor 1910; Barrows 1976; Graham 1979; U.S. Bureau of the Census 1910).

By 1940, the basic unit of enumeration was no longer the family; instead, there were households and quasi-households. A household consisted of the group of persons occupying a group of rooms with either separate cooking equipment or an outside entrance. A single room could qualify as a household only if it had its own cooking facilities or was the only living quarters in the structure. The maximum number of boarders and lodgers in a household was 10; where that number was exceeded, the unit was enumerated as a quasi-household. Quasi-households also included hotels, institutions, military barracks, dormitories, and the like (Jenkins 1987; U.S. Bureau of the Census 1984a).

The procedure in 1950 was similar to that in 1940, with two important exceptions. First, the maximum number of boarders and lodgers in households was reduced from 10 to 4; units with 5 or more boarders and lodgers became quasi-households. Second, students residing at college on Census Day were no longer enumerated at their parental home (U.S. Bureau of the Census 1955, 1984b).

The definition of households was broadened slightly for the census of 1960 to include persons in any single room with direct access to the outside or to a common hallway, whether or not the room had its own cooking facilities. As a result, single rooms in hotels and boardinghouses were more often classified as separate households. The Census Bureau substituted the term *group quarters* for the term *quasi-household,* but the definition remained virtually the same. The 1970 census definitions were almost the

same as those for 1960, except that quarters without direct access to a common hallway were required to have "complete cooking facilities" to qualify as independent households. The definition was tightened further in 1980, as the bureau finally dropped the cooking facilities criterion and all housing units were required to have direct access. Also, the threshold for classification as group quarters was raised from 5 to 10 unrelated individuals, which made them virtually the same as the quasi-households of the 1940 census (U.S. Bureau of the Census 1966, 1976, 1986).

The top two sections of Table 2 summarize the basic changes in census definitions described above. To some extent, the changes cancel one another out. Consider the following three hypothetical cases:

1. A person living in a room with direct access to the outside via a common hallway but without cooking facilities.

2. A person with a room in a house who must pass through the family living quarters to reach the outside but who has a hot plate, sink, and table in the room.

3. A person in a room without direct outside access who shares a kitchen with others but eats in the room.

Case 1 would count as a separate household in 1910, 1960, 1970, and 1980. By contrast, Case 2 would constitute an enumeration unit in all years *except* 1970 and 1980. Case 3 would be listed as a separate unit only in 1880, 1900, and 1910.

The impact of changing definitions depends on the relative frequency of different living arrangements in different periods. Owing to the rising standard of living and technological and architectural changes since the turn of the century, separate entrances and cooking facilities have become more commonplace. In the 1880–1910 census years, both entrances and kitchens were often shared; neither, however, was a requirement for a separate enumeration unit. If an enumerator from the 1980 census could go back in time to 1880 and collect information from the tenements of the Lower East Side, he or she would no doubt find fewer separate units than the enumerators of 1880. Because many "families" lacked direct access to a common hallway, a modern enumerator would find fewer persons living alone and more extended families and secondary families. Paradoxically, however, an enumerator who traveled the opposite direction in time, from 1880 to 1980, and canvassed the condominiums of a southern California strip

Table 2 Summary of major changes in units of sampling and enumeration: Public use samples, 1880–1980

	1880	1900	1910	1940	1950	1960	1970	1980
Available units of enumeration								
Dwelling	X							
"Family" (old definition)	X	X	X					
Household				X	X	X	X	X
Quasi-household				X	X			
Group quarters						X	X	X
Minimum for enumeration as separate household or "family"								
Separate eating table	X	X						
Separate portions of dwelling			X					
Minimal cooking facilities				X	X	X		
Outside entrance				X	X			
Direct access via hallway						X	X	X
Complete cooking facilities							X	
Threshold for sampling unrelated individuals as separate units	30	1[a]	20	5	5	5	5	10
Enumeration of college students at parental home	Y	Y	Y	Y	N	N	N	N

[a]Coresident domestics were included as part of the "family" in the 1900 PUS, but all boarders and lodgers were treated as separate units.

would probably find just about the same number of units as the 1980 enumerator. This is because people who eat separately now tend also to have direct access to a common hallway. In a sense, then, the 1880 census definitions are reasonably compatible with the 1980 census, but the 1980 definitions are incompatible with the 1880 census.

The sorts of living arrangements that fall in the ambiguous region between the different census definitions have become rare. On the whole, one would expect the changes in census definitions of the enumeration unit to have only moderate consequences for the classification of living arrangements. If anything, from 1880 to 1980 the definitions became somewhat more restrictive, meaning that it became more difficult to qualify as a separate unit. These changes could increase the potential for classification as extended families; some extended families that were enumerated in two separate units in 1880 would probably have been counted as a single unit in 1950. By 1980, all units were required to have direct access, which further increases the potential for complex households with extended kin or boarders. The changes in definitions could also have implications for the frequency of primary individuals (heads of household without family), secondary individuals (persons unrelated to the head without family), and secondary family members (persons unrelated to the head with family).[1] An enlargement of enumeration units could increase the proportion of secondary families and secondary individuals and reduce the categories of primary families and primary individuals. Moreover, the adoption of the quasi-household and group quarters classifications since 1940 further reduced the potential number of primary individuals, since by definition primary individuals must be heads of households.

The census data show marked declines in the frequency of extended families and secondary individuals and increases in the proportion of primary families and primary individuals. The decline of secondary families has been so great that since 1970 the Census Bureau has no longer bothered to tabulate them. If the definitions of the enumeration units had remained constant, these changes might have been even greater. Thus, if changing census definitions have had any effect at all, it is probably to understate the extent of change in household structure.

The change in the enumeration of college students alluded to

above would tend to counteract the gradual trend towards definitions encouraging larger and more inclusive households. From 1880 through 1940 college students were counted at their "usual place of abode," which meant that most of the students in dormitories and rooming houses were counted as if they still resided with their parents. Since 1950, such students have been classified as residents of quasi-households, group quarters, or primary or secondary individuals. I have elsewhere described an adjustment procedure to account for the effects of this change (Ruggles 1988: Appendix).

Beyond the formal differences in census definitions across census years, there have also been changes in the enumeration procedures that could have implications for the delineation of enumeration units. There has been a significant improvement in the quality of enumeration since 1940, probably resulting from better recruitment and training of enumerators and increasing efforts to ensure quality control.[2] Thus, one might expect that the formal rules have been more closely followed in recent census years. The adoption of self-enumeration based on forms mailed to the respondent may also have had some effect. As noted above, in 1960 most respondents were mailed forms in advance of the census, to be filled out and later collected by an enumerator. The bureau hoped that self-enumeration would help to reduce enumerator error. In 1970, self-enumeration was taken one step further; in an effort to save money, most people were requested to mail their forms back to the census office. These changes may have contributed to a de facto definition of the enumeration unit as a mailing address, regardless of the formal definition. Although the direction of potential bias is uncertain, self-enumeration may have reinforced a general trend towards more inclusive definitions of households.

The units of analysis available in the public use samples are also affected by the sampling strategies used in their construction. All the public use files incorporate special procedures for persons residing in institutions and large group quarters. Members of large units have been sampled on an individual basis simply by treating each member as if they lived in their own one-person household. This procedure increases the efficiency of the sample by raising the number of observations while still maintaining representativeness.

Unfortunately, since the criteria for designating units to be

sampled on an individual basis have varied, the samples are incompatible for some applications. In the 1980 public use sample, all units with nine or more members unrelated to the householder were classified as group quarters, and members of group quarters were sampled on an individual basis (U.S. Bureau of the Census 1982a). For the public use samples of the period 1960–70, the procedure was similar, except that units with five or more secondary individuals or secondary family members were classified as group quarters and sampled individually (U.S. Bureau of the Census 1972a, 1984a, 1984b). Insofar as it was possible, the creators of the 1940 and 1950 public use samples imposed the 1970 census definitions of households and group quarters. Thus, residents of quasi-households and those in households with five or more unrelated individuals were classified as persons in group quarters and sampled as individuals.

In the 1910 sample up to 20 members of a family could be unrelated to the head before the members were sampled at the individual level (Strong et al. 1989). This higher threshold for individual-level sampling in 1910 allows detailed study of the small boardinghouses that were characteristic of the period. In the case of the 1900 data file, all boarders and lodgers and the institutionalized were sampled as individuals or as secondary families, a strategy that maximized precision at great cost in terms of lost information (Graham 1979). For example, the 1900 system makes it impossible to create an analogue of the group quarters concept used in recent census years, because there is no way to determine the number of persons in the "family" who were unrelated to the head of household.[3]

The 1880 dataset will be a sample of dwellings rather than a sample of families like the census files for 1900 and 1910. This strategy has been adopted because it allows study of the composition of multifamily dwellings and requires only a small compromise of efficiency. Each dwelling will contain one or more families that are closely comparable to the families in the 1900 and 1910 samples. The threshold for individual-level sampling will be 30, which is larger than in any of the previous samples. Thus, all the definitions of group quarters used for the later census years can be reconstructed for 1880 simply by reclassifying family members as members of group quarters, according to the num-

ber of unrelated individuals in the family (Ruggles and Menard forthcoming).

SAMPLE DESIGNS AND TREATMENT OF MISSING DATA

Beyond the differences in the treatment of unrelated groups, there were also differences in the procedures for drawing the census samples. None of the samples is a pure random sample of the population; they all incorporate strategies to enhance representativeness. In the cases of the 1880, 1900, and 1910 data files, the sampling was stratified according to geography by randomly selecting a fixed proportion of households within each microfilm reel or block of census pages (Graham 1979; Strong et al. 1989; Ruggles and Menard forthcoming).

The 1940 file is a systematic sample in which households containing sample lines were selected in inverse proportion to household size. For example, for 1940 every second two-person household containing a sample line was included, and every fifth five-person household with a sample line was included. This procedure accounts for the higher probability of larger households including a sample-line person. However, very large households were oversampled, so a set of weights must be used when processing the 1940 sample. The 1950 sample is similar, except that households with sample-line persons were selected without regard to their size, so analyses of the person records must use weights that are inversely proportional to household size (U.S. Bureau of the Census 1984a, 1984b).

The data files for the three recent census years of 1960 through 1980 are a little different. Because they were produced as by-products of the processing of each census, it was possible to stratify according to characteristics other than geography, such as household type, household size, race, and housing tenure, and this further increases precision (U.S. Bureau of the Census 1972a, 1973, 1982a).

Estimation of sampling error for the various samples is complicated by the differing stratification schemes. It is made even more difficult because of the hierarchical structure of the files. Since census microdata files are cluster samples (ordinarily clustered

by household), standard errors depend on both the number of clusters and the homogeneity of variables within clusters. In the worst case, with perfect homogeneity within clusters, the standard errors for variables would be inversely proportional to the square root of the number of sample units rather than the number of individuals. For variables that are not very homogeneous within clusters, such as age, the relevant number of cases is closer to the total number of persons in the file than to the number of independently selected households (U.S. Bureau of the Census 1972a; Kish 1965).

In practice, few investigators attempt to estimate sample errors even when they are working with only one of these files. The large size of the samples means that in most analyses the standard errors are too small to be of great concern. The effort required to estimate errors for an analysis using the entire series would be so great that it seems improbable that anyone would bother.

Before turning to discussion of specific variables, I should add a word about the treatment of missing, illegible, and inconsistent data. All the census files incorporate some degree of logical editing of missing and inconsistent values. For example, in each file the sex of a wife was assumed to be female. The more common enumerators' errors cannot be resolved through this type of logical computer editing. Thus, the samples of 1880 and 1940–80 incorporate "hot deck" allocation procedures to assign missing values (U.S. Bureau of the Census 1972a; Banister 1980). For each variable, there is a series of criteria for matching a "donor" record used to impute the missing or inconsistent value. These criteria are determined through analysis of the best predictors for each variable. To take an example from the 1940 allocation procedure, if sex was missing or illegible and had not been allocated through logical editing, the sex of the most proximate individual in the file with the same race, age, and marital status was allocated (U.S. Bureau of the Census 1984a). If a perfectly matched donor record could not be found, the record that met the largest number of criteria was used. The donated value was then subjected to consistency checks and rejected if unsuitable. To allow researchers to reconstruct the original data, allocated data items were indicated by a data-quality flag.

The alternative to allocation is simply to exclude cases with missing data from the analysis. In effect, this assumes that indi-

viduals with missing data are representative of the population as a whole. Using allocated data requires the less extreme assumption that persons with missing data are representative of the population that shares their key characteristics, including geographic proximity.

As the discerning reader may have guessed, the specific procedures used to allocate missing and inconsistent data are different in every census year. In a perfect world, the whole thing would be redone on a consistent basis. In practice, however, the differences are relatively insignificant and should not materially affect analysis. A larger problem is that allocation was never carried out for the 1900 and 1910 samples. The Minnesota Social History Research Laboratory is currently developing new versions of these files that incorporate missing-data allocation.

VARIABLES ON HOUSEHOLD COMPOSITION AND DEMOGRAPHY

This and the following sections describe the major changes in the variables included in each public use file. This is intended only as an overview; I describe only variables available before 1960, and even for these variables the discussion is not intended to be comprehensive. Users of the public use samples should pay close attention to the definitions provided with each codebook.

Table 3 summarizes the available variables on household composition. All the samples include a basic variable describing the relationships among the members of the family or household. From 1880 to 1970, the relationship was expressed in reference to a household head. Household headship was defined by the respondents; the only rule was that a married woman residing with her husband could not be reported as head. In 1980, the gender-free concept of "householder" replaced the concept of household head. A householder is defined as the homeowner or leaseholder of the home; if a husband and wife jointly own or lease their home, either may be listed as the householder. The relation-to-householder variable can easily be made compatible with relation-to-head in earlier census years.

The first two rows of Table 3 show the number of categories of household relationship codes available in each public use sample. The earlier census years provide considerably more detail than

Table 3 Summary of available information on household relationships: Public use samples, 1880–1980

	1880	1900	1910	1940	1950	1960	1970	1980
Number of family relationship categories	TBA	35	50	15	15	10	10	14
Number of nonfamily relationship categories	TBA	54	74	7	7	5	5	6
Surname similarity code	Y		Y	Y	Y			
Subfamily relationships	Y	C	C	Y	Y	Y	Y	Y
Secondary family relationships	Y	C	C	Y	Y	Y		
1970 Census Bureau household and family classifications	Y	C[a]	C	Y	Y	Y	Y	C
Shanas family classification	Y	C	C	C	C	C	C	C
Laslett-Hammel classification	Y	C	C	C	C	C[b]	C[b]	C

Note: Blank = variable not available. Y = yes. C = can be constructed. TBA = to be announced.

[a]The 1970 census concept of "group quarters" cannot be precisely constructed for the 1900 PUS.

[b]Cannot be constructed for all households.

the later ones. Partly because of the use of FOSDIC multiple-choice enumeration forms, only 10 family relationships were distinguished in 1960 and 1970. Aunts, uncles, cousins, nephews, nieces, and grandparents of the head were all lumped together in the single category of "other relative." Such kin were quite rare by 1960, but the lack of detail means that detailed household classifications such as the Laslett-Hammel scheme (Laslett 1972) cannot be applied with certainty to all households.

Explicit secondary family relationship codes are provided for the samples of 1880, 1940, 1950, and 1960. For 1900 and 1910, such relationships can generally be inferred from the household relationship codes, which include such categories as boarder's

wife and boarder's child. In addition, the availability of surname or surname similarity codes in 1880, 1910, 1940, and 1950 can often help sort out unclear secondary family relationships.[4] No information on secondary family relationships is available for 1970 or 1980; the Census Bureau decided that secondary families had become so rare that the expense of gathering the information was not justified. As noted earlier, the sampling procedure for 1900 makes it impossible to determine if five secondary individuals are present in a "family," so the 1970 census category of group quarters cannot be precisely replicated, although it can be approximated.

Despite the limitations of the household relationship codes since 1960 and the sampling procedure for secondary individuals in 1900, the public use samples are highly compatible for the study of family and household composition. Virtually all standard household classification systems can be approximated in all census years. Moreover, together with the information on basic demographic characteristics, the household relationship codes provide sufficient information to identify the presence and characteristics of own children, own grandchildren, own parents, own spouses, and own siblings for virtually the entire population in all census years.

The availability of demographic characteristics is shown in Table 4. Age and sex are identical in all census years. The categories of race vary somewhat, but in all census years a basic classification of white, black, American Indian, Chinese, and other can be constructed. Marital status is the same in all census years, except that the "separated" category did not appear until 1950; however, the related category of married–spouse absent exists or can be constructed in all census years.

Information on marital history is more erratic. All census years except for 1880 give sufficient information to determine age at marriage, but in 1900, 1910, and 1950 it is age at last marriage, whereas in 1940 and 1960–80 it is age at first marriage. The two measures are compatible for the subset of the population married only once, which can be identified in every census year from 1910 to 1980. By one means or another, the population married within the past year can be identified for all census years with reasonable consistency. This information is valuable for assessing the changing living arrangements of newlyweds.

Fertility data are widely available. Children-ever-born to ever-

Table 4 Summary of available information on basic demographic characteristics: Public use samples, 1880–1980

	1880	1900	1910	1940	1950	1960	1970	1980
Age	Y	Y	Y	Y	Y	Y	Y	Y
Sex	Y	Y	Y	Y	Y	Y	Y	Y
Race	Y	Y	Y	Y	Y	Y	Y	Y
Marital status[a]	Y	Y	Y	Y	Y	Y	Y	Y
Duration of current marriage		Y	Y		S[b]			
Age at first marriage				S		Y	5	Y
Number of marriages			Y	S[c]	S[c]	Y	5	Y
Married in past year	Y	Y	Y	C,S	C,S	C	C	C
Children ever born		Y	Y	S	S	Y	Y	Y
Children surviving		Y	Y					
Own-child fertility measures	C	C	C	C	C	C	C	C

Note: Blank = variable not available. Y = yes. C = can be constructed. S = sample-line individuals only, 1940 and 1950. 5 = 5% sample only, 1970 PUS.

[a]The "separated" category of marital status is not available before 1950; however, the similar category of married, spouse absent, can be constructed for all census years.

[b]Duration of current marital status.

[c]The 1940 and 1950 censuses indicated whether married more than once.

married women is given in every sample since 1900. Even more important, children present in the household can be linked to their mothers in every census year, allowing analysis of the timing of fertility change by own-child techniques. Differential mortality can be estimated for 1900 and 1910 by using the variable on children surviving; for other periods, two-census methods can be applied to estimate mortality for certain population subgroups.

GEOGRAPHIC CODES AND VARIABLES ON ETHNICITY AND MIGRATION

The geographic codes are the most frustrating ones. Precise information on locality was gathered in every census year, but because of privacy regulations this information has been omitted from the public use samples of the period 1940–80. The samples provide

Table 5 Summary of available information on geography: Public use samples, 1880–1980 (selected variables)

	1880	1900	1910	1940	1950	1960	1970	1980
Smallest geographic area identified (thousands)				100	100	250	250	100
State	Y	Y	Y	Y	Y	Y	N,ST	Y
Urban/rural residence	Y	Y	Y			Y[a]	N,ST	c
Farm identifier[b]	C	Y	Y	Y	Y	Y	Y	Y
Identification of large central cities by name	Y	Y	Y	Y	Y		SM	a,b
County urban population	Y	Y	Y					
SMA				Y	Y			
SMSA							SM	a,b
County or county group	Y	Y	Y	Y	Y		SM	a,b

Note: Blank = variable not available. C = can be constructed. Y = variable available. N = neighborhood characteristics samples, 1970 PUS. ST = state samples, 1970 PUS. SM = SMSA samples, 1970 PUS. a = A sample, 1980 PUMS. b = B sample, 1980 PUMS. c = C sample, 1980 PUMS.

[a]Not available for all states.

[b]Definition of farm varies.

different and usually incompatible geographic identifiers for each census year, and for 1970 and 1980 the Census Bureau created three versions of each sample with alternate geographic variables. The variables that can be made roughly compatible across multiple census years are shown in Table 5. The first row of the table gives the minimum size of identified geographic units allowed in each census year under the privacy rules.

The greatest difficulty for most historical applications is that there are no direct measures of rural/urban residence for 1940 or 1950. Persons residing outside of standard metropolitan areas can be identified, but this includes persons residing in cities that were large by the standards of 1880 or 1900. The 1940 and 1950 samples also include a variable called state economic area, which identifies county groups with homogeneous economies. We can easily classify as "rural" county groups with more than a given

percentage of rural occupations, such as farming and forestry. However, this definition cannot be applied to the 1960 sample, where no county group data are available, and can only be roughly approximated for 1970 and 1980, where different county groupings are given.

The rural/urban classification given in 1970 and 1980 is a simple dichotomy based on residence either in "urbanized areas" or places with 2,500 or more population. The urbanized-area census concept cannot be precisely replicated for the 1880–1910 samples, but it can be approximated. Consistent identification of metropolitan areas is difficult. The 1940–50 definitions of standard metropolitan area are similar to the 1970–80 standard metropolitan statistical areas, although there were subtle changes in the criteria in virtually every decade. However, these modern definitions are partly based on commuting ties and measures of metropolitan character that are not available in the earlier census years. The samples for 1880 and 1900 do provide the size of the urban population for the county and adjacent counties, and this can be used to create a crude analogue of metropolitan area. The same variable can be constructed for 1910 by linking the sample to the county data file prepared by the Inter-University Consortium for Political and Social Research (ICPSR No. 003).

A variety of particular places can be identified across all census years. With inconsequential exceptions, state of residence is indicated for the entire population in all census years. Sixty-one of the largest cities can be identified across all census years except 1960, although in many instances the boundaries of those cities have changed during the past century. In addition, the state economic areas of 1940–50 and the county groups of 1970 and 1980 can be reconstructed for 1880–1910, but again there have been some changes in county boundaries.

The public use samples are a rich source of information on immigration and ethnicity (Table 6). Birthplace is available for all census years, and parental country of birth is available for every year except 1980, so both immigrants and their children are generally identifiable. The country codes are given with striking detail in all the samples. Since the map of central and eastern Europe has been twice redrawn in the twentieth century, consistent identification of national origin for that region is difficult. For students of central and eastern European immigration, however, national boundaries are a poor guide to ethnicity in any

Table 6 Summary of available information on ethnic origins and migration: Public use samples, 1880–1980 (selected variables)

	1880	1900	1910	1940	1950	1960	1970	1980
Birthplace (country, state)	Y	Y	Y	Y	Y	Y	Y	Y
Citizenship/ naturalization		Y	Y	Y	Y		5	Y
Parental birthplaces (country)	Y	Y	Y	S	S	Y	15	
Parental birthplaces (state)	Y	Y	Y	S	S			
Residence five years earlier				Y		Y	15	Y
Year of immigration		Y	Y				5	Y
Mother tongue			Y	S	S	Y	15	Y
Speaks English		Y	Y					Y
Spanish surname	Y			Y	Y	Y	Y	Y

Note: Blank = variable not available. Y = variable available. S = sample-line individuals only, 1940 and 1950. 5 = 5% sample only, 1970 PUS. 15 = 15% sample only, 1970 PUS.

period. Most will make better use of the mother-tongue codes, which are available from 1910 to 1980. Year of immigration is a critical variable for studies of assimilation; unfortunately, it is given only in the samples for 1900, 1910, 1970, and 1980. Despite this limitation, year of immigration allows direct comparison of the assimilation of the "new" immigrants of the early twentieth century with the assimilation of the "new" immigrants of the late twentieth century.

There are also several indicators of internal migration available across census years. State of birth is given in all the samples. In addition, the samples from 1880 to 1950 provide parental state of birth, and the 1940, 1960, 1970, and 1980 samples show place of residence five years earlier, subject to the limited precision dictated by the privacy rules.

VARIABLES ON ECONOMIC STATUS AND EMPLOYMENT

When economists or demographers learn that no information on income was gathered before 1940, they are generally dismayed.

Table 7 Summary of available information on economic status and employment: Public use samples, 1880–1980 (selected variables)

	1880	1900	1910	1940	1950	1960	1970	1980
Wage and salary income				Y	Y	Y	Y	Y
Total income					Y	Y	Y	Y
Occupation (see text)	Y	Y	Y	Y	Y	Y	Y	Y
Industry	C	C	Y	Y	Y	Y	Y	Y
Home ownership		Y	Y	Y		Y	Y	Y
Mortgaged		Y	Y					Y
Rent/home value				Y		Y	Y	Y
Domestics in household	C	C	C	C	C	C	C	C
Employment status (employer, self-employed, wage or salary worker)			Y	Y	Y	Y	Y	Y
Period worked in census year				Y	S	Y	Y	Y
Hours worked previous week				Y	Y	Y	Y	Y
Period unemployed in census year	Y	Y	Y					Y
Currently unemployed			Y	Y	Y	Y	Y	Y

Note: Blank = variable not available. Y = variable available. C = can be constructed. S = sample-line individuals only, 1940 and 1950.

Even in 1940, the income variable is of limited use, since the census reported only wage and salary income. This means that income in 1940 cannot be effectively used for persons such as farmers, doctors, or shopkeepers. Thus, income serves as the key indicator of economic status only for the last 30 years of the public use series (Table 7).

Occupation is the main variable on economic status available for the public use samples of 1880–1910. The use of information on occupation to classify people according to economic status is full of potential pitfalls. Many occupational titles are too vague to provide a clear indication of economic status. Moreover, because

of economic, social, and even linguistic changes, the occupational hierarchy is constantly shifting. Indeed, some major occupations at the turn of the century, such as copyist, have virtually ceased to exist, and there are new occupations, like airplane pilots and computer operators.

The use of occupation as an indicator of economic status is further complicated by the lack of an occupational classification system oriented to this purpose. The Census Bureau classifications are measures of type of work as much as they are of economic rank. Thus, for example, under the 1950 occupational classification system, stockbrokers were grouped together with newsboys and "hucksters and peddlers," under the subcategory of "sales workers."

Among the general classification systems the census has produced during the past century, the 1950 system is easiest to replicate in all census years. From 1950 through 1970, the Census Bureau had a reasonably consistent system of 11 broad occupational categories divided into several hundred specific occupational codes; the specific changes in this period are described in U.S. Bureau of the Census 1968 and 1972b. The nine-category system of 1940 can easily be reconciled with that of 1950, since the specific codes used in the public use samples are virtually the same. There was a major revision in 1980, but the Census Bureau has provided the necessary information to optimize backwards compatibility (U.S. Bureau of the Census 1989). The 1900 public use sample employs the 1950 classification system as well as that for 1900, and the 1910 sample gives both the 1980 and the 1910 classifications. The 1880 public use sample will take advantage of the painstaking efforts of Ann Miller and others, who worked on the occupational recodes for 1910 by adapting the 1910 occupational data dictionaries to meet the needs of the 1880 sample.

The 1950 occupational codes can be reordered to conform more closely to our intuitive notions of economic rank, but any such system would necessarily be somewhat arbitrary. The Social History Research Laboratory has developed an alternate strategy. About 200 specific occupations or narrow occupational groupings can be consistently identified across all census years. It is a simple matter to calculate median and standard deviation of income for each of these categories in the 1950 public use sample.

The information on income is then used to construct two indexes: first, an economic score, reflecting the relative income of each category in 1950, and second, a precision score, which can be used to weed out those jobs for which the title provides a poor predictor of income. The final step is to attach the two indexes to all eligible persons in each census year. To help account for the decline in status associated with the feminization of occupations like clerical work, the indexes are calculated separately for each gender whenever the specific occupational title includes sufficient cases to allow it. The economic score should not be viewed as a true proxy for income; it is simply a more subtle occupational classification than those provided by the Census Bureau.

Like any occupational classification system, this one cannot control for changes in the occupational hierarchy. We can assess the effects of such changes from 1950 on simply by tracing the correlation between economic score and income, but there is no effective means of estimating the reliability of the index in the early period. The precision of the economic score can be expected to decline as one gets farther away from 1950. In practice, the economic scores are most useful when grouped into quintiles or deciles to avoid false precision.

A second problem is that the economic score is a poor proxy for the income of farmers, who constituted the single largest occupational category in the early public use samples. We can get some idea of the economic status of farmers by assuming homogeneity within counties and attaching the value of farm and value of farm product variables from the county-level data files created by the Inter-University Consortium for Political and Social Research. In the long run, the individual-level agricultural schedules from 1880 will be linked to the public use sample. This will give precise measures of farm size in that year and a means of assessing the usefulness of the county-level variables in 1900 and 1910. None of these variables, of course, will be comparable to information in the later public use samples.

Other indicators of economic status are scarce. Home ownership is available for all samples except 1880 and 1960, and mortgage information is given in 1900, 1910, and 1980. Rent and home value appear in 1940 and 1960–80. For the early period, the number of domestic servants is a useful means of identifying the wealthy, but domestic service declined dramatically after

1910. Each of these measures has limitations as an indicator of economic status, but taken together, they can serve as a useful check on the occupational information.

Occupation must also serve as the main indicator of labor-force participation in the early period. Prior to 1940, the census asked about usual occupation, whether or not the person was actually employed at the time of the census. It was not until 1940 that the census added questions on hours worked the previous week and weeks worked the previous year. There is some controversy about the reliability of occupation as a measure of labor-force participation in the period from 1880 to 1910. The disagreement stems from the sharp rise in apparent labor-force participation between 1900 and 1910, especially for women. Some of the early analysts felt that the 1910 census had substantially overcounted the participation of women. Now, most scholars argue that the 1910 census figures are substantially correct, and that the problem lies with an undercount in earlier census years (U.S. Bureau of the Census 1943; Jaffe 1956; Openheimer 1970; Conk 1980, 1981). The latter explanation seems more likely. The Census Bureau made a special effort to gather comprehensive information on occupation in 1910. The enumerators' instructions specified that "an entry should be made in this column for *every* person enumerated. The occupation, if any, followed by a woman, or a child, of any age, is just as important, for census purposes, as the occupation followed by a man. Therefore it must never be taken for granted, without inquiry, that a woman, or child, has no occupation" (U.S. Bureau of the Census 1910: 32). In cases where the respondent was neither employed nor temporarily unemployed, the enumerator was instructed to enter "own income" or "none" in the occupation column. By contrast, in the 1900 census occupation was to be reported only for persons over the age of 10 who were employed. Moreover, in 1910 the census asked additional questions on whether one was "employer, employee, or working on own account" or currently out of work. Even though these questions appeared after the occupational inquiry on the census form, they may have helped to screen the employed population. All things considered, it seems most plausible that the 1910 enumeration was reasonably accurate with regard to labor-force participation, and that the 1880 and 1900 figures were undercounted.

From 1940 on, the bureau asked questions on hours worked

the previous week and weeks worked the previous year to determine labor-force participation, and the results are probably more reliable. The same questions served as screening devices for the occupational inquiry, so there is a close correspondence in the public use samples among these variables.

Many historians are interested in class distinctions measured according to workers' relationship to the means of production. The questions on employment status and class of worker that the census has asked since 1910 addresses this issue directly by classifying all workers as employers, self-employed, or employees. Highly detailed information on occupational titles is available for 1880, 1900, and 1910; since occupational title is a reliable predictor of employment status for most titles, the 1910 census can be used to impute employment status in the earlier census years.

Finally, all the public use samples incorporate at least one measure of unemployment. From 1880 to 1910, the census asked how long each worker had been unemployed during the census year. Estimates of unemployment rates based on these data are unrealistically low; therefore, the unemployment variable is generally regarded as suspect. However, the unemployment information is still useful if the goal is not to construct unemployment rates but rather to analyze the characteristics of the population affected by unemployment. From 1910 on, the Census Bureau asked increasingly detailed questions about current unemployment, and the reliability of the questions has greatly improved.

VARIABLES ON EDUCATION, LITERACY, AND OTHER TOPICS

Users of the recent public use samples have come to depend almost as much on years of schooling as they do on income. Unfortunately, as shown in Table 8, this variable is not available in the early public use samples. Mean years of schooling for population subgroups can, however, be inferred by using the school attendance question, which appears in all census years. The method is analogous to the singulate mean age at marriage.[5] Historians have used school attendance creatively, but because it is only a characteristic of children, it is not nearly as powerful as educational attainment.

Variables on literacy are available for 1880 through 1910.

Table 8 Summary of available information on education, literacy, veteran status, and disabilities: Public use samples, 1880–1980 (selected variables)

	1880	1900	1910	1940	1950	1960	1970	1980
School enrollment	Y	Y	Y	Y	S	Y	Y	Y
Can read	Y	Y	Y					
Can write	Y	Y	Y					
Years of schooling				Y	S	Y	Y	Y
Veteran status			Y[a]	S	S	Y	15	Y
Disability (definition varies)	Y		Y				5	Y

Note: Blank = variable not available. Y = variable available. S = sample-line individuals only, 1940 and 1950. 5 = 5% sample only, 1970 PUS. 15 = 15% sample only, 1970 PUS.
[a]Civil war veterans only.

Although illiteracy was doubtless underreported, it has nevertheless proven to be a valuable inquiry. From 1940 on, the question on years of schooling completed has been essentially unchanged. The final two variables in Table 8 are veteran status, available in limited form since 1910, and disability, which the census asked in rather different forms at the beginning and end of the series.

The public use samples include a substantial number of variables that I have not discussed. The 1880 census, for example, included a detailed question on current morbidity, and the 1910 census asked about mother's mother tongue and father's mother tongue. The variables I have omitted are either given only in a single census year or have been asked only since 1960, so they are not relevant to a discussion of the long-range comparability of the public use samples.

CONCLUSION

In a discussion of comparability issues, the differences among the samples necessarily receive more ink than their similarities. The scope of the census has greatly increased during the past century, and several important census definitions have changed. But it still remains recognizably the same thing. In the areas of household and family structure, demographic behavior, and immigration, all the census years are closely comparable, so investigators need not

fear pooling the data from different census years for multivariate analyses of change. Working with the geographic codes requires patience and occasional compromises, but for most applications acceptably comparable geographic categories can be constructed. The early censuses are especially limited when it comes to analyses of economics and education. Even in these areas, however, investigators can do a lot if they count things creatively.

One area of incompatibility I have not stressed is the differences in the layout and numerical coding of each public use sample. Even in cases where the variables are simple and virtually identical, such as sex or marital status, the different samples often use different coding schemes. Moreover, although all the samples are arranged in column format, the basic variables are located in different columns for each census year. An exception to these problems exists for 1960 and 1970. When the Census Bureau released a new version of the 1960 sample in conjunction with the 1970 sample, the numerical coding and variable locations were made compatible.

For the past several years, the Social History Research Laboratory at the University of Minnesota has been constructing compatible-format versions of all the public use samples. These files incorporate a variety of compatible constructed variables, including household and family structure classifications, characteristics of own parents and spouses, pointers to location within the household of parents and spouses, and the basic variables needed for own-child fertility analysis.

To date, the Social History Research Laboratory samples have been limited to internal use for research and teaching by Minnesota faculty and graduate students. The documentation of the samples is as yet inadequate for public release, and the data format and coding schemes need to be refined further. Common-format data files will probably be released through ICPSR within the next two years. These files should greatly simplify analysis and encourage the use of the census files as an integrated data series.

NOTES

1 These are the 1940–70 PUS definitions of primary and secondary individuals; in the 1900 and 1910 codebooks, the term *primary individual* is equivalent to the term *secondary individual* in standard usage.

Table 9 Estimates of net percentage census undercount

Census year	Coale-Zelnik-Rives estimates[a]	Census Bureau estimates
1880	6.5	
1890	7.2	
1900	6.8	
1910	6.5	
1920	6.8	
1930	5.3	
1940	5.0	5.6
1950	3.5	4.4
1960		3.3
1970		2.9
1980		1.4

Sources: Coale and Zelnik 1963: 181–82; Coale and Rives 1973; Census Bureau estimates; Fay et al. 1988.
[a]Whites and blacks only.

2 On the recruitment of enumerators and efforts to ensure quality control, see Anderson 1988. Estimates of net underenumeration vary widely; see, for example, U.S. Bureau of the Census 1982b. According to the most widely accepted figures, overall net undercount declined from 5.6% in 1940 to 1.4% in 1980 (Siegel 1974; Fay et al. 1988; cf. Land et al. 1984; Siegel 1968; Coale 1955; Price 1947). For the period before 1950, the estimates of underenumeration are problematic because of the weakness of the vital statistics and the lack of postenumeration surveys. Francis Walker and Carroll Wright, the late nineteenth-century directors of the census, both claimed that net underenumeration was under 1% around the turn of the century (U.S. Bureau of the Census 1916: 16), but such a figure cannot be believed. Coale and Zelnik (1963) and Coale and Rives (1973) have estimated net undercount for blacks and whites in the period 1880–1950 by the birth reconstruction method, and these figures are given in Table 9. It appears that most of the improvement in coverage has occurred since the 1940 census. The greatest reduction in undercount has occurred among whites. Indeed, the estimates reported by Coale and Rives (ibid.) suggest that the undercount of black males was actually more severe in 1970 than in 1880.

Historians have frequently expressed concern about underreporting in the census (e.g., Sharpless and Shortridge 1975). In comparison with alternative cross-sectional sources, however, the census looks pretty good. We can be reasonably confident that the response rate was better than 90% in all the census years for which we have public use samples, and this figure compares favorably with the best of recent social surveys. For the nineteenth century, no alternative data source even comes close to the census in terms of coverage.

3 Although the 1900 public use sample does not provide sufficient information to adopt precisely consistent criteria for households and group quarters, the

categories of the 1940–70 public use samples can be roughly approximated. The 1900 household record indicates whether or not 10 or more boarders were resident in the enumeration unit. Thus, one can classify as residents of group quarters persons in units with 10 or more boarders, military installations, boarding schools, college dormitories, old age homes, poorhouses, convents, homes for unwed mothers, and institutions. All other individuals in the 1900 census should be considered to reside in households. By these criteria, the 1900 definition of group quarters is a subset of the 1940–70 public use sample definition, but the difference is small.

4 The same codes can be useful for interpreting ambiguous family relationships, such as "brother-in-law," which could mean either sister's husband or wife's brother.

5 For an example of the method, see Stevens forthcoming.

REFERENCES

Anderson, Margo J. (1988) The American Census: A Social History. New Haven, CT: Yale University Press.

Banister, Judith (1980) "Use and abuse of census editing and imputation." Asian and Pacific Census Forum 6: 1–20.

Barrows, R. G. (1976) "Instructions to enumerators for completing the 1900 census population schedule." Historical Methods 9: 201–12.

Coale, Ansley J. (1955) "The population of the United States in 1950 classified by age, sex, and color: A revision of the figures." Journal of the American Statistical Association 50: 16–54.

———, and N. W. Rives (1973) "A statistical reconstruction of the black population of the United States, 1880–1970: Estimates of true numbers by age and sex, birth rates, and total fertility." Population Index 39: 3–36.

Coale, Ansley J., and Melvin Zelnik (1963) New Estimates of Fertility and Population in the United States. Princeton, NJ: Princeton University Press.

Conk, Margo A. (1980) The United States Census and Labor Force Change. Ann Arbor: UMI Research Press.

——— (1981) "Accuracy, efficiency, and bias: The interpretation of women's work in the U.S. Census of Occupation, 1890–1940." Historical Methods 14: 65–72.

Department of Commerce and Labor (1910) Annual Reports for 1909. Washington, DC: U.S. Government Printing Office.

Eckler, A. Ross (1972) The Bureau of the Census. New York: Praeger.

Fay, Robert E., Jeffrey S. Passel, J. G. Robinson, and Charles D. Cowan (1988) 1980 Census of Population and Housing. Evaluation and Research Reports. The Coverage of Population in the 1980 Census.

Graham, Stephen N. (1979) 1900 Public Use Sample: User's Handbook. Seattle: University of Washington, Center for Demography and Ecology.

Jaffe, A. J. (1956) "Trends in the participation of women in the working force." Monthly Labor Review 79: 559–65.

Jenkins, Robert (1987) Procedural History of the 1940 Census of Population and Housing. Madison: University of Wisconsin Press.

Kish, Leslie (1965) Survey Sampling. New York: Wiley.
Land, Kenneth C., George C. Hough, and Marilyn M. McMillan (1984) "New midyear age-sex-color specific estimates of the U.S. population for the 1940s and 1950s: Including a revision of coverage estimates for the 1940 and 1950 censuses." Demography 21: 623–45.
Laslett, Peter (1972) "Introduction," in Peter Laslett and Richard Wall (eds.) Household and Family in Past Time. Cambridge: Cambridge University Press: 1–65.
Openheimer, Valerie K. (1970) The Female Labor Force in the United States: Demographic and Economic Factors Governing Its Growth and Changing Composition. Population Monograph No. 5. University of California, Berkeley.
Price, Daniel O. (1947) "A check on underenumeration in the 1940 census." American Sociological Review 12: 44–49.
Ruggles, Steven (1988) "The demography of the unrelated individual, 1900–1950." Demography 25: 521–36.
———, and Russell R. Menard (forthcoming) "A public use sample of the 1880 Census of Population." Historical Methods.
Sharpless, John B., and Ray M. Shortridge (1975) "Biased underenumeration in census manuscripts: Methodological implications." Journal of Urban History 1: 409–39.
Siegel, Jacob S. (1968) "Completeness of coverage of the nonwhite population," in David M. Heer (ed.) Social Statistics and the City. Cambridge, MA: MIT and Harvard University, Joint Center for Urban Studies: 13–54.
——— (1974) "Estimates of the coverage of the population by age, sex, and race in the 1970 census." Demography 11: 1–23.
Stevens, David (forthcoming) "Life-course transitions to adulthood, 1900–1970." Journal of Family History.
Strong, Michael A., Samuel H. Preston, Ann R. Miller, Mark Hereward, Harold R. Lentzner, Jeffrey R. Seaman, and Henry C. Williams (1989) User's Guide: Public Use Sample, 1910 Census of Population. Philadelphia: University of Pennsylvania, Population Studies Center.
U.S. Bureau of the Census (1910) Thirteenth Census of the United States: Instructions to Enumerators. Washington, DC: U.S. Government Printing Office.
——— (1916) Special Reports of the Twelfth Census. Supplementary Analysis and Derivative Tables. Washington, DC: U.S. Government Printing Office.
——— (1943) Sixteenth Census of the United States. Population. Comparative Occupation Statistics for the United States, 1870 to 1940, by Alba M. Edwards. Washington, DC: U.S. Government Printing Office.
——— (1955) The 1950 Censuses: How They Were Taken. Washington, DC: U.S. Government Printing Office.
——— (1966) The 1960 Censuses of Population and Housing: Procedural History. Washington, DC: U.S. Government Printing Office.
——— (1968) Changes between the 1950 and 1960 Occupation and Industry Classifications, by John A. Priebe. Technical Paper 18. Washington, DC: U.S. Government Printing Office.
——— (1972a) Public Use Samples of Basic Records from the 1970 Census: De-

scription and Technical Documentation. Washington, DC: U.S. Government Printing Office.

——— (1972b) 1970 Occupation and Industry Classifications in Terms of Their 1960 Occupation and Industry Elements, by John A. Priebe, Joan Heinkel, and Stanley Greene. Technical Paper 26. Washington, DC: U.S. Government Printing Office.

——— (1973) Technical Documentation for the 1960 Public Use Sample. Washington, DC: U.S. Government Printing Office.

——— (1976) U.S. Census of Population and Housing: 1970 Procedural History. Washington, DC: U.S. Government Printing Office.

——— (1979) Twenty Censuses—Population and Housing Questions: 1790–1980. Washington, DC: U.S. Government Printing Office.

——— (1982a) Public Use Samples of Basic Records from the 1980 Census: Description and Technical Documentation. Washington, DC: U.S. Government Printing Office.

——— (1982b) "Coverage of the national population in the 1980 census by age, sex, and race: Preliminary estimates by demographic analysis." Current Population Reports, Series P-23, No. 115. Washington, DC: U.S. Government Printing Office.

——— (1984a) Census of Population, 1940: Public Use Sample Technical Documentation. Washington, DC: U.S. Government Printing Office.

——— (1984b) Census of Population, 1950: Public Use Sample Technical Documentation. Washington, DC: U.S. Government Printing Office.

——— (1986) Census of Population and Housing (1980): History, 1980 Census of Population and Housing. Washington, DC: U.S. Government Printing Office.

——— (1989) "The relationship between the 1970 and 1980 industry and occupation classification systems." Technical Paper No. 59. Washington, DC: U.S. Government Printing Office.

U.S. Census Office (1882) Report of the Superintendent of the Census (November 1, 1881). Washington, DC: U.S. Government Printing Office.

——— (1883) Tenth Census of the United States: 1880. Statistics of Population. Washington, DC: U.S. Government Printing Office.

——— (1895) Eleventh Census of the United States: 1890. Population, pt. 1. Washington, DC: U.S. Government Printing Office.

Walker, Francis A. (1888) "The eleventh census of the United States." Quarterly Journal of Economics 2: 136–61.

Wright, Carroll, and William C. Hunt (1900) The History and Growth of the United States Census. Washington, DC: U.S. Government Printing Office.