INTRODUCTION

The Minnesota Historical Census Projects

Steven Ruggles and Russell R. Menard

The census is our fundamental source of information about American social structure in the past. No other source can compete with respect to population coverage and reliability.[1] For the period before the mid-twentieth century, the census provides the only data on population characteristics that are not profoundly distorted with respect to class, race, gender, or education. Moreover, the census is the only historical source that provides comprehensive geographic coverage and broad chronological scope. Indeed, many Americans in earlier generations left no surviving trace of their existence except for their listings in the census. Thus, we must frequently turn to census data for basic historical generalizations about the lives of ordinary people.

Quantitative studies of long-term social change have always relied on published tabulations of the census. The published data are invaluable, but they do have significant limitations. Until recently, the Census Bureau itself carried out only the most rudimentary analyses with the data it collected. The topics addressed by census publications have always focused on contemporary concerns, concerns that have shifted dramatically over the past two centuries. For example, although the Census Bureau has gathered data on family and household composition since 1880, it was not until 1940 that census publications began to provide statistics in this area. On the other hand, the Progressive Era censuses provided great detail on nativity and ethnicity, reflecting the contemporary concern with immigration. The high costs of tabulation before the introduction of modern data-processing equipment meant that few cross-classifications of census data were feasible, and the Census Office had to make difficult choices about what to measure. As a result, much information collected by the census was never tabulated at all. Thus, census publications do not exploit the full potential of the census enumerations.

Development of Census Microdata Files

In the 1960s, the "new social historians" adopted an innovation that multiplied the uses of census data. Instead of using the aggregate published tabulations of the census, scholars like Stephan Thernstrom (1963) turned directly to the enumerators' manuscripts. These manuscripts provided information on individual households or persons—what we now call *microdata*, to distinguish it from data describing population aggregates.

The microdata collected by enumerators contain far richer information than was ever published in the census volumes. The new social historians found that by using microdata they could make tabulations tailored to their particular research questions. In fact, the nineteenth-century manuscript censuses allowed researchers to address issues never contemplated by the Census Office (as the Census Bureau was known before 1902). This was especially true for the census years from 1850 onward, when the enumerators' manuscripts began to provide information on individuals as well as households. By combining the characteristics of different individuals within the same household, historians could construct new categories of information. For example, they analyzed the occupations of spouses and residence with extended kin, even though such information was collected only inadvertently by the census and was never tabulated. Moreover, historians using the nineteenth-century census manuscripts soon began to use analytic methods that did not even exist when the censuses were first tabulated, from multiple regression to own-child fertility analysis.

At about the same time that historians began using nineteenth-century census microdata, the Census Bureau released the first national census microdata. Social science was booming in the early 1960s, and researchers flooded the Census Bureau with requests for specialized tabulations on particular topics. Some social scientists were even beginning to use computers. The Bureau came up with a seemingly obvious but innovative solution: it created a 1-in-1,000 extract of the basic microdata it was using to create the tabulations for the published census volumes and released the extract in machine-readable form to researchers who could then make their own tables (U.S. Bureau of the Census 1963). The data file was called a Public Use Sample, to distinguish it from the census samples being used internally by the Bureau. To preserve confidentiality, the

Census Bureau removed the names, addresses, and other potentially identifying information.

The 1960 Public Use Sample was an immediate success. It allowed researchers to create their own measures and to adopt the latest multivariate analytic tools; but the sample did have two significant limitations. First, the sample size was relatively small. The 1-in-1,000 sample density yielded information on about 180,000 persons. Given the modest capacity of computers in 1964, this was a lot of cases; but as researchers began to use the sample for detailed analysis of small population subgroups, its limitations became apparent. Halliman Winsborough, one of the early users of the 1960 sample, later recalled "analyses evaporating in a welter of empty cells." Second, the 1960 Public Use Sample provided limited geographic information. In its zeal to preserve confidentiality, the Census Bureau stripped off all information on places below the state level. This meant, for example, that it was impossible to extract a subsample of the New York City population.

The Census Bureau addressed both problems with the release of the 1970 Public Use Samples (U.S. Bureau of the Census 1972a). The total sample size rose some sixty-five-fold, to 12 million persons. In addition, the 1970 samples provided a variety of alternate geographic codes, although the Census Bureau still did not identify any places with a population of less than 250,000.

In conjunction with the 1970 Public Use Samples, the Census Bureau released an expanded version of the 1960 Public Use Sample, enlarging the sample density from 1-in-1,000 to 1-in-100 (U.S. Bureau of the Census 1973). The record layout and coding schemes of the new 1960 sample were reorganized to be virtually identical to those of the 1970 samples. This compatibility made it easy for investigators to pool data from 1960 and 1970 and thereby incorporate chronological change into their analyses.

By the late 1970s, the public use samples from the 1960 and 1970 censuses had become essential tools of American social science. It was in this climate that two separate teams of researchers independently came up with the idea of extending the series backward by creating historical public use samples for earlier census years. Samuel Preston directed projects at the University of Washington and the University of Pennsylvania to produce a 1-in-750 sample of the 1900 census and a 1-in-250 sample of the 1910 census (Graham 1980; Strong et al. 1989). Meanwhile, Winsborough and a group of other demographers at the University of Wisconsin in collaboration with the Census Bureau created 1-in-100 samples for the census years 1940 and 1950 (U.S. Bureau of the Census 1984a, 1984b). Both Preston and Winsborough suffered federal budget cuts and ran short of money.

All four historical census files were eventually completed, but not without strain. In the mid-1980s, the future looked dim for historical public use samples of the census. The 1900, 1940, and 1950 samples had all become available by 1984, but they failed to attract the scholarly interest that had been envisioned by their creators. The 1910 sample, which did eventually stimulate a great deal of research, was far behind schedule. Both Preston and Winsborough decided they had done enough.

Undaunted by numerous warnings about the perils of data-collection projects, at the Social History Research Laboratory of the University of Minnesota we decided to pick up where Preston and Winsborough left off. The 1880 census was the obvious candidate for the next national microdata project. It was the first census to include such key inquiries as marital status, family relationships, and parental birthplaces, but because the Census Office ran out of money in 1881, these variables were never tabulated. In late 1988, we submitted a modest proposal to the National Institutes of Health to create a 1-in-500 sample of the 1880 census. The study group reviewed the proposal with enthusiasm, but it wanted one small modification: it requested an enlargement of sample size to allow study of small population subgroups. We complied, raising the proposed sample density from 1-in-500 to 1-in-100. The proposal was funded, and in the summer of 1989 we began to enter data on some 550,000 individuals (Ruggles and Menard 1990). We finished the 1880 Public Use Microdata Sample (PUMS) in 1994.

Thus encouraged, we decided to complete the series of national census microdata files. In 1992, we began work on a 1-in-100 sample of the 1850 population census, with funding from the National Science Foundation. Because 1850 was the first census year to gather information at the level of individuals, it is the logical starting point for the series of national census microdata. The final version of the 1850 sample was completed in 1994.

We recently received funding for a 1-in-100 sample of the 1920 census from the National Institutes of Health. To protect the privacy of census respondents, census enumeration manuscripts remain confidential for seventy-two years. The public release of the 1920 manuscripts in 1992 made a national sample for that year feasible.[2] The 1920 project, which involves data entry of information on 1.1 million persons, will be completed in 1998.

National PUMS files therefore exist or are in preparation for all but four census years since 1850. The 1890 manuscript census was destroyed in a fire, so there is no possibility of creating a national census sample in that year. The remaining census years are 1860, 1870, and 1930. We are presently working on proposals to create new samples for the 1860 and 1870 census years. The 1930 census manuscripts will not be released until 2002, so that sample will be delayed.

In the meantime, we hope to enlarge and enhance the census files for 1900 and 1910. The 1900 sample was the first of the historical files to be produced, and at 100,000 persons it is also the smallest by a wide margin. Moreover, the sample design for 1900 is somewhat incompatible with later census years. We therefore plan to apply for funds to

enlarge the 1900 sample to approximately 760,000 cases. The 1910 sample is larger than the 1900 sample, and it has already been expanded by adding an oversample of the black population. In addition, we are presently working with Myron Gutmann of the University of Texas to add an oversample of the Hispanic population in 1910.

As historical work was extending the series of national census files backward in time, the Census Bureau was extending it forward by releasing machine-readable samples of the 1980 and 1990 censuses (U.S. Bureau of the Census 1983, 1993). We therefore now have a series of public use microdata samples of the U.S. census covering the years 1850, 1880, 1900, 1910, 1920, 1940, 1950, 1960, 1970, 1980, and 1990. The general characteristics of these samples are shown in table 1. If we can obtain funds to fill the remaining gaps, we will eventually have a large national sample for every census year from 1850 onward except for 1890.

Strengths and Weaknesses of the Public Use Microdata Samples

The range of potential topics that can be addressed with the national census files includes household composition, fertility, life-course transitions, ethnicity, immigration, internal migration, female labor-force participation, the household economy, industrial and occupational structure, urbanization, nuptiality, and education.[3] Compared with the community-level census microdata files created by Thernstrom and many other historians in the 1960s and 1970s, the national census microdata samples do have some liabilities. Many community-level samples created by historians linked local census data to other local sources, such as tax lists or business directories, thereby enriching the basic census data. Moreover, in some cases historians linked individuals from census year to census year, thus allowing longitudinal analysis. The national census microdata files do not incorporate such enrichment: they are independent, random cross-sectional samples for successive years without any individual-level information added from other sources.[4]

The national census files do, however, offer three key strengths that make them far more powerful than the census samples of particular communities ordinarily used by historians. These strengths are complete geographic coverage, large scale, and broad chronological scope.

Complete geographic coverage is important not only because it allows generalization at the national level. Paradoxically, one of the greatest advantages of national samples is their potential for studying the ways in which local conditions affect behavior. Such analysis has always been one of the goals of community studies, and the authors of

TABLE 1
Characteristics of the Public Use Files, 1850–1990

Census year	Principal investigator/ sample version	Release date		Sample density	No. of cases (in thousands)		No. of variables[a]		
		Preliminary	Final		Households	Persons	Household record	Person record	Sample line
1850	Menard	1993	1994	1 in 100	38	200	27	35	
1880	Ruggles	1991	1994	1 in 100	107	502	35	42	
1900	Preston		1980	1 in 760	27	100	70	33	
1910	Preston		1989	1 in 250	89	366	30	40	
1920	Ruggles	1995	1998	1 in 100	243	1,057	50	50	
1940	Winsborough		1984	1 in 100	391	1,352	30	55	27
1950	Winsborough		1984	1 in 100	461	1,922	23	39	42
1960	Census Bureau	1964	1971	1 in 100	579	1,800	63	53	
1970	Census Bureau								
	15% state sample		1972	1 in 100	744	2,030	63	62	
	15% county sample		1972	1 in 100	744	2,030	60	62	
	15% neighborhood		1972	1 in 100	744	2,030	61	62	
	5% state sample		1972	1 in 100	744	2,030	75	59	
	5% county sample		1972	1 in 100	744	2,030	72	59	
	5% neighborhood		1972	1 in 100	744	2,030	73	59	
1980	Census Bureau								
	A sample		1983	1 in 20	4,500	10,812	67	78	
	B sample		1983	1 in 100	900	2,162	67	78	
	C sample		1983	1 in 100	900	2,162	67	78	
1990	Census Bureau								
	5% sample		1992	1 in 20	5,528	12,500	78	79	
	1% sample		1992	1 in 100	1,106	2,500	78	79	

[a]Excluding data-quality flags.

such studies have frequently criticized national analyses because they obscure local and regional diversity. Ironically, individual community studies do not permit the study of the impact of local conditions or geographic diversity: these topics can only be addressed by comparing localities. Comparison of community studies is complicated by inevitable differences in measures and methods among different historians.

The national public use census files allow systematic comparisons across space. City, county, and even ward-level local characteristics on topics such as agriculture, manufacturing, religion, voting, and property taxes are readily available, for the most part in machine-readable form. We can easily link these local characteristics to the historical census files for the period 1850 through 1920. We can then carry out contextual analyses of the effects of local conditions on individual and family behavior. Although such contextual analysis is still in its infancy, the early studies are indeed impressive (Elman 1993, Landale and Tolnay 1991; also see Ruggles forthcoming).[5]

The second strength of the national public use census files is their large size. The number of cases available for each census year ranges from the hundreds of thousands to the tens of millions. This allows study of small and geographically dispersed population subgroups. For example, University of Minnesota graduate students using the historical public use samples have examined topics such as the professionalization of nursing; American Indian fertility patterns; the living arrangements of elderly urban blacks; the demography of the prison population; the gender composition of clerical workers; and the living arrangements of parentless children. These research topics could not have been pursued using a general social survey of the scale ordinarily undertaken by academic social scientists. Indeed, even the largest social survey carried out by the government—the Current Population Survey—is far too small for the detailed analysis of topics such as American Indian fertility or the composition of the clerical work force. The public use samples are the only general source of microdata available for any period with sufficient cases to study such small population subgroups.

The third—and most important—strength of the historical public use census files is their potential for the study of social and economic change over long periods of time. There is no other consistent source of information about the American population spanning more than a few decades. Despite frequent changes in subject content and modifications of enumeration procedures, the core of the census has remained remarkably stable over the past century and a half. So far, however, few researchers have exploited the great potential of the national census files for the study of long-term change. Instead, most investigators use the samples as isolated cross-sections. At Minnesota we recently began to compile a bibliography of research using the PUMS; to date, over 80 percent of the studies use only one of the eleven census years currently available.

The national census microdata files have not been widely used to study change mainly because it is difficult to use more than one national census file at a time. Each sample has a different format, different coding schemes, and different documentation. Since the original 1960 and 1970 samples, at least five separate research teams have been involved in the creation of the samples, and each has had its own ideas on how to organize the data. We are faced with eight different occupational classifications with a total of thirty-two hundred different categories and at least seven incompatible classifications for variables such as birthplace, household relationship, and institution type.

Documentation for the eleven existing samples is contained in eleven separate volumes totaling over three thousand pages. These volumes are organized differently from one another, and their treatment of comparability issues is often cursory. The 1960 and 1970 Public Use Samples constitute the only exception to the general rule of incompatibility. It has been relatively easy to use the 1960 and 1970 census years in combination; as a result, most of the research using more than one public use microdata sample has focused on these two census years.

The incompatibility of the PUMS in their present form means that multisample studies require a large initial investment of time and money to prepare the data. To use multiple census years in the same analysis, each investigator must prepare a special-purpose compatible extract of selected variables. This ad hoc approach has led to duplication of effort among the few investigators using multiple census years. Moreover, because of the complexity of the census files and the often subtle differences among them, the potential for error is large.

To reduce the incompatibility problems among the national census files, the Social History Research Laboratory at the University of Minnesota is now converting the series of public use samples into a single coherent form with uniform documentation. This project, funded by the National Science Foundation, is called the Integrated Public Use Microdata Series (IPUMS).

The IPUMS promises to be a powerful tool for the study of social change. The database will include information on over 50 million individuals spread over 140 years of extraordinary social and economic change. We expect that the unprecedented potential to locate individual behavior in time and spatial context will generate important new research on topics such as fertility, urbanization, immigration, household composition, and occupational structure.

Organization of This Issue

This issue of *Historical Methods* describes the historical census projects at the University of Minnesota. The first section, by Diana L. Magnuson and Miriam L. King, focus-

es on the quality and compatibility of the source material. The census is an artifact of the past and, like all other historical sources, must be evaluated in the context of the purposes and procedures of its creators. Accordingly, Magnuson and King turn to contemporary sources to describe how and why changes in the subject content of the census occurred and the history of enumeration procedures of the population censuses since 1850.

The second section describes the making of the IPUMS. It includes short articles by the staff of the IPUMS project on sample designs and sampling errors, record layout and coding procedures, the special problems posed by occupational and socioeconomic coding, and the creation of variables on family interrelationships. The IPUMS is still a work in progress, and it is likely that we will make further changes in its design. We solicit suggestions for improvements from potential users; communications should be sent to ruggles@atlas.socsci.umn.edu.

Finally, in the third section, we turn to the creation of the historical census samples for 1850, 1880, and 1920. The articles in this section—written by the programming, research, and data-entry staffs of the three projects—briefly cover data-entry software, data-entry procedures, data cleaning and error control, and the creation of data dictionaries, with special attention to the very complex occupational and geographic dictionaries. We hope this section will be useful to investigators contemplating similar projects.

NOTES

This research was funded by National Institutes of Health grants 5R01-HD25839-04, 1R01-HD29015-01A1, and 1R01-HD32325 and National Science Foundation grants SBR-9118299-02 and SBR-9210903-01. The contributions to this theme issue were edited with the assistance of Matthew Mulcahy, David Ryden, and Beth Salerno of the University of Minnesota. We are also grateful for the help of Barbara Kahn of Heldref Publications.

1. It has long been fashionable among historians to question the reliability, accuracy, and representativeness of the census (see, for example, Hill 1993; Sharpless and Shortridge 1975). In fact, the older censuses compare favorably with modern social survey data in all these respects. Many recent sociological surveys have nonresponse rates far greater than those of the nineteenth-century censuses. For a recent discussion of census coverage issues, see King and Magnuson (1995).
2. The 1940 and 1950 census years are still covered by census confidentiality rules, so those historical samples were created by Census Bureau employees at far greater cost. A preliminary 1-in-1,000 subsample of the 1920 PUMS is now available; the final 1-in-100 version will be complete in 1998.
3. For a discussion of some applications of the historical census microdata samples, see Ruggles (1993).
4. Record linkage—either between census years or between census records and other sources—is always a difficult and painstaking undertaking, but in the case of the national census files it is either impossible or extremely expensive. For the period from 1940 onward, no record linkage can be carried out because those census years are still protected by confidentiality rules and therefore provide no information that can be used to identify individual cases. Although some linkage is theoretically possible in the earlier years, the costs would be high. Most community-level census files created by historians include every case from a particular locality, making record linkage fairly straightforward. By contrast, the national census files are low-density random samples, so the identification of links is many times slower.
5. From 1940 onward, the potential for contextual analysis of the census files will be somewhat more limited; because of the census confidentiality rules, the public use files for those years do not identify geographic location with great precision.

PART 1. Historical Comparability of the U.S. Census

Who and What Determined the Content of the U.S. Population Schedule Over Time

Diana L. Magnuson

The population census is often described as a mirror of society, reflecting characteristics of the U.S. population at a specific point in time. While that is unarguably true, the census is more than simply a mirror of society in a "snapshot" sense. The census as a document is an image of the ideas and issues that were most important to American society at the time of each enumeration and, more important, a reflection of the people and groups who influenced reform of each decennial census. The actions and views of strategically placed individuals and interest groups were extremely significant in determining what questions appeared on the census schedules and how those questions were worded. While the historical and political climate indirectly shaped the census content, it would be a mistake to see the population schedule as a pure reflection of general societal trends or to analyze the schedule purely as a text, without considering the views and intentions of those who designed the questionnaire. This article analyzes which individuals and groups most affected the population schedule from 1850 on and, in relation to these actors, discusses some notable changes in the scope and content of the population schedule.[1]

Table 1 shows the availability of variables in each census year; table 2 presents the number of changed questions from year to year. The two tables are not strictly compatible, however. Table 1 refers to variables in the machine-readable data, while table 2 relates to questions actually appearing on the census schedules. A cursory examination of the Integrated Public Use Microdata Series (IPUMS) availability table reveals questions that persist from census to census, other queries that appear and drop out across census years, and still others that make a single appearance. The construction of IPUMS explicitly highlights the persistence or nonpersistence of population schedule questions over time. Why were these changes made? Who or what influenced change in census content? This article explores these issues with illustrative examples and does not comprehensively treat all population schedule questions.

Throughout its history, the population schedule has been marked by seasons of reform (e.g., 1850, 1880, 1940, and 1960) followed by periods of quiescence or retreat (see figures 1–4). Particularly significant was the change between 1840 and 1850, marking the end of one era of census taking and the beginning of another, persisting to today. Under the old system (1790–1840), the household was the enumeration unit. Asking how many people with a given set of characteristics (e.g., white, male, and over 16) were found in a household, the enumerator supplied the appropriate tallies in separate columns devoted to each specified set of characteristics. Thus, if the population were simultaneously broken down into two sexes, two races, and three age periods, twelve columns would be needed to record all the requisite information. As census users demanded coverage of more subjects and more detailed classifications, this household-based schedule format became increasingly awkward. With a total of eighty columns (see figure 1), the 1840 population schedule demonstrated the practical limits to this enumeration style.

Beginning in 1850, individuals rather than households were the enumeration unit, with the same questions put to every person separately (see figure 2). Thus, data on sex, race, occupation, and age could be collected via only four columns, with no predetermined limit placed on the number of occupation and age subcategories. This change was rev-

TABLE 1
Variable Availability, 1850-1990

Variable Availability Key:

X = Available for all cases in this year
5 = Available in the 5 percent sample in 1970
15 = Available in the 15 percent sample in 1970
S = Available for sample-line persons in 1940 or 1950
C = Variable constructed in this year
. = Variable not available in this year

	1850	1880	1900	1910	1920	1940	1950	1960	1970	1980	1990
HOUSEHOLD RECORD:											
TECHNICAL:											
Record type	X	X	X	X	X	X	X	X	X	X	X
Census year	X	X	X	X	X	X	X	X	X	X	X
Household serial number	X	X	X	X	X	X	X	X	X	X	X
Data set number	X	X	X	X	X	X	X	X	X	X	X
Number of person records following	X	X	X	X	X	X	X	X	X	X	X
Dwelling size	X	X	.	.	X	.	.	.	.	.	.
Number of households in dwelling	X	X	.	.	X	.	.	.	.	.	.
Household weight	.	.	.	.	.	X	X	.	.	.	.
Self-weighting sample identifier	.	.	.	.	.	X	X	.	.	.	.
Sample line person number	.	.	.	.	.	X	X	.	.	.	.
Subsample number	C	X	C	C	C	X	X	X	X	X	X
Report form	.	.	.	.	.	.	.	X	.	.	.
Sample identifier	.	.	.	.	.	.	.	.	.	.	X
Microfilm locator variables	X	X	X	X	X	.	.	.	.	.	.
Enumeration date	X	X	.	.	X	.	.	.	.	.	.
Date of receipt	.	X	.	.	.	.	.	.	.	.	.
LOCATION CHARACTERISTICS: *											
Region	X	X	X	X	X	X	X	X	X	X	X
State	X	X	X	X	X	X	X	X	X	X	X
City identifier	X	X	X	X	X	X	X	.	.	X	X
Ward	.	.	X	X	X	.	.	.	.	.	.
Urban/rural status	X	X	X	X	X	.	.	X	X	X	X
City population	X	X	X	X	X	X	X	.	.	X	X
County	X	X	X	X	X	.	.	.	.	.	.
County group	.	.	.	.	.	.	.	.	X	X	.
PUMA	.	.	.	.	.	.	.	.	.	.	X
State economic area	C	C	C	C	C	X	X	.	.	.	.
Size of place	X	X	X	X	X	C	C	.	.	C	C
SMA/SMSA/MSA	C	C	C	C	C	X	X	.	X	X	X
Central city	C	C	C	C	C	X	X	X	X	X	X
Metro status	C	C	C	C	C	X	X	X	X	X	X
Urbanized area	.	.	.	.	.	.	.	.	.	X	.
Size of urbanized area, 1970	.	.	.	.	.	.	.	.	X	.	.
Area identified, 1970	.	.	.	.	.	.	.	.	X	.	.
Geographic division, 1970	.	.	.	.	.	.	.	.	X	.	.

	1850	1880	1900	1910	1920	1940	1950	1960	1970	1980	1990
HOUSEHOLD RECORD (CONT.)											
Planning service area	.	.	.	.	.	.	.	.	.	.	X
Enumeration district	.	X	X	.	X	.	.	.	.	.	.
Enumeration district match ID	.	.	.	.	.	X	X	.	.	.	.
Address	.	X	.	.	X	.	.	.	.	.	.
GROUP QUARTERS:											
Group quarters type	X	X	X	X	X	X	X	X	X	X	X
Institution type	X	X	X	X	X	X	X	X	X	X	X
Institution funding code	X	X	C	X	X	C	C	C	C	C	C
DWELLING PHYSICAL CHARACTERISTICS:											
Access to unit	.	.	.	.	.	.	.	X	X	X	.
Kitchen or cooking facilities	.	.	.	.	.	.	.	X	X	X	X
Number of rooms	.	.	.	.	.	.	.	X	X	X	X
Plumbing facilities	.	.	.	.	.	.	.	X	X	X	X
Hot and cold water	.	.	.	.	.	.	.	X	X	.	.
Bathtub or shower	.	.	.	.	.	.	.	X	X	.	.
Flush toilet	.	.	.	.	.	.	.	X	X	.	.
Basement	.	.	.	.	.	.	.	X	X	.	.
Year structure built	.	.	.	.	.	.	.	X	X	X	X
Condition of housing	.	.	.	.	.	.	.	X	.	.	.
Units at address	.	.	.	.	.	.	.	.	X	X	.
Units in structure	.	.	.	.	.	.	.	X	X	X	X
Source of water	.	.	.	.	.	.	.	X	15	X	X
Sewage	.	.	.	.	.	.	.	X	15	X	X
Number of bathrooms	.	.	.	.	.	.	.	X	15	X	.
Bedrooms	.	.	.	.	.	.	.	X	5	X	X
Stories	.	.	.	.	.	.	.	X	5	X	.
Elevator	.	.	.	.	.	.	.	X	5	X	.
APPLIANCES, MECHANICALS, ETC:											
Telephone	.	.	.	.	.	.	.	X	X	X	X
Television	.	.	.	.	.	.	.	X	5	.	.
UHF television	.	.	.	.	.	.	.	.	5	.	.
Radio	.	.	.	.	.	.	.	X	5	.	.
Washing machine	.	.	.	.	.	.	.	X	5	.	.
Clothes dryer	.	.	.	.	.	.	.	X	5	.	.
Dishwasher	.	.	.	.	.	.	.	.	5	.	.
Home food freezer	.	.	.	.	.	.	.	X	5	.	.
Air conditioning	.	.	.	.	.	.	.	X	15	X	.
Heating equipment	.	.	.	.	.	.	.	X	X	X	.
ECONOMIC:											
Farm	C	C	X	X	X	X	X	X	X	X	X
Farm schedule	.	.	X	X	X	.	.	.	.	.	.
Ownership of dwelling	.	.	X	X	.	X	.	X	X	X	X
Mortgage status	.	.	X	X	.	.	.	.	.	X	X
Second mortgage	.	.	.	.	.	.	.	.	.	X	X
Value of dwelling	.	.	.	.	.	X	.	X	X	X	.
Commercial use	.	.	.	.	.	.		X	X	X	X
Rent includes farming land	.	.	.	.	.	.	.	X	.	.	.
Sales of farm products	.	.	.	.	.	.	.	X	X	X	X
Acreage of property	.	.	.	.	.	.	.	X	X	X	X

	1850	1880	1900	1910	1920	1940	1950	1960	1970	1980	1990
HOUSEHOLD RECORD (CONT.)											
House acreage	.	.	.	.	.	.	X	.	.	X	.
Mortgage payment	.	.	.	.	.	.	.	.	.	X	X
Second mortgage payment	.	.	.	.	.	.	.	.	.	.	X
Real estate taxes in payment	.	.	.	.	.	.	.	.	.	X	X
Insurance premiums in payment	.	.	.	.	.	.	.	.	.	X	X
Property taxes and insurance cost	.	.	.	.	.	.	.	.	.	X	.
Insurance cost	.	.	.	.	.	.	.	.	.	.	X
Property taxes	.	.	.	.	.	.	.	.	.	.	X
Rent	.	.	.	.	.	X	.	X	X	X	X
Gross rent	.	.	.	.	.	.	.	X	X	X	X
Meals included in rent	.	.	.	.	.	.	.	.	.	.	X
Condominium fee	.	.	.	.	.	.	.	.	.	.	X
Mobile home costs	.	.	.	.	.	.	.	.	.	.	X
Electricity cost	.	.	.	.	.	.	.	.	X	X	X
Gas cost	.	.	.	.	.	.	.	.	X	X	X
Water cost	.	.	.	.	.	.	.	.	X	X	X
Cost of heating fuel	.	.	.	.	.	.	.	.	X	X	X
Total family income	.	.	.	.	.	.	S	X	X	X	X
Total family wage/salary income	.	.	.	.	.	X	.	.	.	.	.
Total houshold income	.	.	.	.	.	.	.	.	.	X	X
VACANCY:											
Vacancy status	.	.	.	.	.	.	.	.	X	X	X
Usual home elsewhere	.	.	.	.	.	.	.	.	.	X	X
Boarded-up status	.	.	.	.	.	.	.	.	.	X	X
Duration of vacancy	.	.	.	.	.	.	.	.	X	X	X
Second home	.	.	.	.	.	.	.	.	5	.	.
OTHER VARIABLES:											
Condominium status	.	.	.	.	.	.	.	.	X	X	X
Cooking fuel	.	.	.	.	.	.	.	.	5	X	.
House heating fuel	.	.	.	.	.	.	.	X	5	X	X
Water heating fuel	.	.	.	.	.	.	.	.	5	X	.
Automobiles available	.	.	.	.	.	.	.	X	15	X	.
Trucks and vans available	.	.	.	.	.	.	.	.	.	X	.
Vehicles available	.	.	.	.	.	.	.	.	.	C	X
When moved in	.	.	.	.	.	.	.	.	.	X	X
Verified data record	X	X	.	.	X	X	X	.	.	.	.
Respondent's relationship to head	.	.	.	.	.	X	.	.	.	.	.
Specified rent unit	.	.	.	.	.	.	.	.	.	.	X
Specified value unit	.	.	.	.	.	.	.	.	.	.	X
Linquistic isolation	.	.	.	.	.	.	.	.	.	.	X
PERSON RECORD:											
TECHNICAL:											
Record type	X	X	X	X	X	X	X	X	X	X	X
Census year	X	X	X	X	X	X	X	X	X	X	X
Household serial number	X	X	X	X	X	X	X	X	X	X	X
Data set number	X	X	X	X	X	X	X	X	X	X	X
Person number in unit	X	X	X	X	X	X	X	X	X	X	X
Sample line weight	.	.	.	.	.	X	X	.	.	.	.

	1850	1880	1900	1910	1920	1940	1950	1960	1970	1980	1990
PERSON RECORD (CONT.)											
Self-weighting sample line person	.	.	.	.	.	X	X	.	.	.	.
Sample line record	.	.	.	.	.	X	X	.	.	.	.
CORE DEMOGRAPHIC:											
Relationship	.	X	X	X	X	X	X	X	X	X	X
Age	X	X	X	X	X	X	X	X	X	X	X
Sex	X	X	X	X	X	X	X	X	X	X	X
Race	X	X	X	X	X	X	X	X	X	X	X
Marital status	.	X	X	X	X	X	X	X	X	X	X
Age at first marriage	.	.	.	.	.	S	.	X	5	X	.
Duration of current marital status	.	.	X	X	.	.	S	.	.	.	.
Times married	.	.	.	X	.	S	S	X	5	X	.
Children ever born	.	.	X	X	.	S	S	X	X	X	X
Children surviving	.	.	X	X	.	.	.	.	.	.	.
SELECT CONSTRUCTED VARIABLES:											
Mother's location in household	C	C	C	C	C	C	C	C	C	C	C
Probable step/adopted mother	.	C	C	C	C	C	C	C	C	C	C
Mother's linking rule	.	C	C	C	C	C	C	C	C	C	C
Father's location in household	C	C	C	C	C	C	C	C	C	C	C
Father's linking rule	.	C	C	C	C	C	C	C	C	C	C
Number of own family members	.	C	C	C	C	C	C	C	C	C	C
Number of own children	.	C	C	C	C	C	C	C	C	C	C
Number of own children under age 5	.	C	C	C	C	C	C	C	C	C	C
Spouse's location in household	C	C	C	C	C	C	C	C	C	C	C
Spouse's linking rule	.	C	C	C	C	C	C	C	C	C	C
ETHNICITY/NATIVITY:											
Birthplace	.	X	X	X	X	X	X	X	X	X	X
Mother's birthplace	.	X	X	X	X	S	S	X	15	.	.
Father's birthplace	.	X	X	X	X	S	S	X	15	.	.
Generation	.	C	C	C	C	C	C	C	C	.	.
Ancestry	.	.	.	.	.	.	.	.	.	X	X
Citizenship	.	.	X	X	X	X	X	.	5	X	X
Year naturalized	.	.	.	.	X	.	.	.	.	.	.
Year of immigration	.	.	X	X	X	.	.	.	5	X	X
Years in United States	.	.	X	.	.	.	.	.	.	.	.
Mother tongue	.	.	.	X	X	S	.	X	15	.	.
Mother's mother tongue	.	.	.	X	X	.	.	.	.	.	.
Father's mother tongue	.	.	.	X	X	.	.	.	.	.	.
Language spoken	.	.	.	X	.	.	.	.	.	X	X
Speaks English	.	.	X	X	X	.	.	.	.	X	X
Ability to speak English	.	.	.	.	.	.	.	.	.	X	X
Mother tongue, 1940	.	.	.	.	.	S	.	.	.	.	.
Puerto Rican stock	.	.	.	.	.	.	.	X	15	.	.
Hispanic origin	.	.	.	.	.	.	.	.	5	X	X
Spanish surname	C	C	.	.	C	X	X	X	X	X	.
Spanish surname, 1950	.	.	.	.	.	.	X	.	.	.	.
EDUCATION:											
School attendance	X	X	X	X	X	X	S	X	15	X	X
Months in school	.	.	X	.	.	.	.	.	.	.	.
Highest grade completed	.	.	.	.	.	X	S	X	X	X	.

	1850	1880	1900	1910	1920	1940	1950	1960	1970	1980	1990
PERSON RECORD (CONT.)											
Educational attainment, 1990	.	.	.	.	.	.	.	.	.	.	X
Highest grade attended	.	.	.	.	.	.	S	.	.	X	.
Finished highest grade	.	.	.	.	.	.	S	.	.	X	.
Type of school	.	.	.	.	.	.	.	X	15	X	X
Vocational training	.	.	.	.	.	.	.	.	5	.	.
Field of vocational training	.	.	.	.	.	.	.	.	5	.	.
Literacy	X	X	X	X	X	.	.	.	.	.	.
WORK:											
Employment status	.	.	.	.	.	X	X	X	X	X	X
Occupation, 1950	X	X	X	C	X	C	X	C	C	C	C
Occupation	X	X	X	X	X	X	X	X	X	X	X
Industry, 1950	.	.	C	C	X	C	X	C	C	C	C
Industry	.	.	.	X	X	X	X	X	X	X	X
Class of worker	.	.	.	X	X	X	X	X	X	X	X
Occupation, 1980	.	.	.	X	.	.	.	.	.	.	.
Industry, 1980	.	.	.	X	.	.	.	.	.	.	.
Occupation code, 1920	.	.	.	.	X	.	.	.	.	.	.
Weeks worked last year	.	.	.	.	.	X	S	X	X	X	X
Hours worked last week	.	.	.	.	.	X	X	.	.	X	X
Usual hours worked	.	.	.	.	.	.	.	.	.	X	X
Months unemployed	.	X	X	.	.	.	.	.	.	.	.
Weeks unemployed	.	.	.	X	X	.	.	.	.	X	.
Continuous weeks unemployed	.	.	.	.	.	X	S	.	.	.	.
Year last worked	.	.	.	.	.	.	.	X	X	X	X
Out of work April 15	.	.	.	X	.	.	.	.	.	.	.
Main activity last week	.	.	.	.	.	.	X	.	.	.	.
Worked last week	.	.	.	.	.	.	.	.	.	.	X
Worked last week, 1950	.	.	.	.	.	.	X	.	.	.	.
Absent from work last week	.	.	.	.	.	.	.	.	.	X	X
Looking for work	.	.	.	.	.	.	X	.	.	X	X
Available for work	.	.	.	.	.	.	.	.	.	X	X
Had job last week	.	.	.	.	.	.	X	.	.	.	.
Worked last year	.	.	.	.	.	.	.	X	X	X	X
Usual occupation	.	.	.	.	.	S	.	.	.	.	.
Usual industry	.	.	.	.	.	S	.	.	.	.	.
Usual class of worker	.	.	.	.	.	S	.	.	.	.	.
Occupation, labor reserve	.	.	.	.	.	.	S	.	.	.	.
Industry, labor reserve	.	.	.	.	.	.	S	.	.	.	.
Class of worker, reserve	.	.	.	.	.	.	S	.	.	.	.
INCOME:											
Total personal income	.	.	.	.	.	.	S	X	X	X	X
Wage and salary income	.	.	.	.	.	X	S	X	X	X	X
Business and farm income	.	.	.	.	.	.	S	X	.	.	.
Non-farm business income	.	.	.	.	.	.	.	.	X	X	X
Farm income	.	.	.	.	.	.	.	.	X	X	X
Nonwage/salary income over $50	.	.	.	.	.	X	.	.	.	.	.
Social security income	.	.	.	.	.	.	.	.	X	X	X
Welfare income	.	.	.	.	.	.	.	.	X	X	X
Interest, dividend, rental income	.	.	.	.	.	.	.	.	.	X	X

	1850	1880	1900	1910	1920	1940	1950	1960	1970	1980	1990
PERSON RECORD (CONT.)											
Retirement income	.	.	.	.	.	.	.	.	.	.	X
Other income	.	.	.	.	.	.	S	X	X	X	X
Real estate value	X	.	.	.	.	.	.	.	.	.	.
Wage/salary of rest of family	.	.	.	.	.	.	S	.	.	.	.
Business income of rest of family	.	.	.	.	.	.	S	.	.	.	.
Other income of rest of family	.	.	.	.	.	.	S	.	.	.	.
Deductions for retirement	.	.	.	.	.	S	.	.	.	.	.
Poverty status	.	.	.	.	.	.	.	.	X	X	X
Poverty cutoff	.	.	.	.	.	.	.	.	X	.	.
MIGRATION: **											
Migration status	.	.	.	.	.	X	S	X	15	X	X
State 1 or 5 years ago	.	.	.	.	.	X	S	.	X	X	X
County group 5 years ago	.	.	.	.	.	.	.	.	.	X	.
Country 5 years ago	.	.	.	.	.	X	S	.	15	.	X
SMA/SMSA 1 or 5 years ago	.	.	.	.	.	X	S	.	.	X	X
Central city 5 years ago	.	.	.	.	.	X	.	.	.	X	X
SEA 1 or 5 years ago	.	.	.	.	.	X	S	.	.	.	.
Migration PUMA	.	.	.	.	.	.	.	.	.	.	X
Metro status 1 or 5 years ago	.	.	.	.	.	X	S	X	15	.	.
Farm status 1 or 5 years ago	.	.	.	.	.	X	S	.	.	.	.
Same incorp place 5 years ago	.	.	.	.	.	X	.	.	.	.	.
Distance between counties	.	.	.	.	.	X	S	.	.	.	.
Year moved into unit	.	.	.	.	.	.	.	X	15	.	.
Migration weight	.	.	.	.	.	.	.	.	.	X	.
ACTIVITY FIVE YEARS AGO:											
Working 5 years ago	.	.	.	.	.	.	.	.	X	X	.
Armed forces 5 years ago	.	.	.	.	.	.	.	.	X	X	.
College 5 years ago	.	.	.	.	.	.	.	.	X	X	.
Occupation 5 years ago	.	.	.	.	.	.	.	.	5	.	.
Industry 5 years ago	.	.	.	.	.	.	.	.	5	.	.
Class of worker 5 years ago	.	.	.	.	.	.	.	.	5	.	.
DISABILITY:											
Disability limiting work	.	.	.	.	.	.	.	.	5	X	X
Disability preventing work	.	.	.	.	.	.	.	.	5	X	X
Duration of disability	.	.	.	.	.	.	.	.	5	.	.
Public transportation disability	.	.	.	.	.	.	.	.	.	X	.
Mobility limitation	.	.	.	.	.	.	.	.	.	.	X
Personal care limitation	.	.	.	.	.	.	.	.	.	.	X
Sickness	.	X	.	.	.	.	.	.	.	.	.
Blind	X	X	.	X	.	.	.	.	.	.	.
Deaf and dumb	X	X	.	X	.	.	.	.	.	.	.
Idiotic	X	X	.	.	.	.	.	.	.	.	.
Insane	X	X	.	.	.	.	.	.	.	.	.
Maimed	.	X	.	.	.	.	.	.	.	.	.
VETERAN STATUS:											
Veteran status	.	.	.	.	.	S	S	X	15	X	X
Veteran status, 1940	.	.	.	.	.	S	.	.	.	.	.
Period of veteran's service, 1940	.	.	.	.	.	S	.	.	.	.	.
Military service since 1980	.	.	.	.	.	.	.	.	.	.	X

	1850	1880	1900	1910	1920	1940	1950	1960	1970	1980	1990
PERSON RECORD (CONT.)											
Military service 1975-1980	.	.	.	.	.	.	.	.	.	X	X
Military service during Vietnam	.	.	.	.	.	.	.	.	15	X	X
Military service 1955-1964	.	.	.	.	.	.	.	.	.	X	X
Military service during Korea	.	.	.	.	.	.	.	X	15	X	X
Military service during WWII	.	.	.	.	.	.	S	X	15	X	X
Military service during WWI	.	.	.	.	.	S	S	X	15	X	.
Military service any other time	.	.	.	.	.	.	S	X	15	X	X
Civil War veteran	.	.	.	X	.	.	.	.	.	.	.
Mortality status of veteran father	.	.	.	.	.	S	.	.	.	.	.
Years of active service	.	.	.	.	.	.	.	.	.	.	X
WORK LOCATION:											
Place of work: SMSA	.	.	.	.	.	.	.	X	15	X	.
Place of work: state	.	.	.	.	.	.	.	X	15	X	X
Place of work: county group	.	.	.	.	.	.	.	.	.	X	.
Place of work: central city	.	.	.	.	.	.	.	.	.	X	.
Place of work: place size	.	.	.	.	.	.	.	.	.	X	.
Place of work: PUMA	.	.	.	.	.	.	.	.	.	.	X
Transportation to work	.	.	.	.	.	.	.	X	15	X	X
Carpooling	.	.	.	.	.	.	.	.	.	X	.
Vehicle occupancy	.	.	.	.	.	.	.	.	.	X	X
Travel time to work	.	.	.	.	.	.	.	.	.	X	X
Time of departure for work	.	.	.	.	.	.	.	.	.	.	X
OTHER VARIABLES:											
Surname similarity	X	X	.	X	X	X	X	.	.	.	.
Name	X	X	.	.	X	.	.	.	.	.	.
Subfamily number	.	C	C	C	C	C	C	C	C	C	C
Subfamily relationship	.	C	C	C	C	C	C	C	C	C	C
Type of subfamily	.	C	C	C	C	C	C	C	C	C	C
Subfamily size	.	C	C	C	C	C	C	C	C	C	C
Age in months	X	X	.	X	.	X	.	.	.	.	.
Month of birth	.	X	X	.	.	.	X	.	.	.	.
Quarter of birth	.	.	.	.	.	.	.	X	X	X	.
Year of birth	.	.	X	.	.	.	.	.	.	.	.
Married within year	X	X	.	.	.	.	.	.	.	.	.
Quarter of first marriage	.	.	.	.	.	.	.	X	5	X	.
Marriage ended by death	.	.	.	.	.	.	.	.	5	X	.
Other infirmity/misfortune	X	.	.	.	.	.	.	.	.	.	.
Crime	X	.	.	.	.	.	.	.	.	.	.
Social security enrollment	.	.	.	.	.	S	.	.	.	.	.

NOTES:

* Availability of geographic variables varies by sample version for 1970-1990.

** 1950 gives migration with respect to 1 year ago, while other censuses use 5 years ago.

Data quality flags are excluded from the table.

SOURCES:

Graham 1980; Ruggles et al. 1993, 1994; Strong et al. 1989; U.S. Bureau of the Census 1972a, 1973, 1982, 1984a, 1984b, 1992.

TABLE 2
Changes in Population Schedule Questions: Additions and Deletions by Census Year

Year	No. of questions Added	Removed	Total
1850			13
1860	1	0	14
1870	7	1	20
1880	11	5	26
1890	21	12	35
1900	7	14	28
1910	8	4	32
1920	6	9	29
1930	11	8	32
1940	28	10	50
1950	21	23	38
1960	23	26	35
1970	20	14	41
1980	8	16	33
1990	3	3	33

Sources: Wright and Hunt (1900, *The History and Growth of the United States Census* (Washington: GPO), U.S. Bureau of the Census (1989), *200 Years of U.S. Census Taking: Population and Housing Questions, 1790–1990* (Washington: GPO).

olutionary in its implications. More substantive areas could be covered without sacrificing legibility; answers could be recorded using detailed classifications (e.g., ages 0 to 100 plus); aggregate results could be reported in tables cross-tabulating any two variables; and the burden of tabulating the returns shifted to a large clerical staff at the central Census Office.[2]

The unwieldy nature of the 1840 population schedule led to numerous inaccuracies in the enumeration, prompting Congress to consider changing the population schedule.[3] In response to complaints from statisticians both in and out of Congress about the accuracy of the 1840 count, Congress appointed a Census Board to investigate the form and content of the census.[4] This was the first time in the history of the census that decisions surrounding the decennial census were put in the hands of men specifically appointed to wrestle with such issues.

The Census Board proposed six separate schedules for the 1850 enumeration.[5] Schedule No. 1, the population schedule, included thirteen questions: dwelling house number; family number; name of every person; age; sex; color; profession, occupation or trade of each male over 15 years of age; value of real estate owned; place of birth; married within the year; attended school within the year; persons over 20 years of age who cannot read and write; whether deaf and dumb, blind, insane, idiotic, pauper, or convict. The Board then presented its work to Congress, fully expecting the schedules to be rubberstamped. Congress, however, had different ideas. Unwilling to relinquish entirely its authority over the census, Congress examined the Census Board's schedules and contributed its own input.[6]

The final make-up of the 1850 schedules highlights the sectional conflict that influenced schedule content. For decades, the North and South had clashed over the proper scope of federal versus state government. Given the dispute over slavery, the South wanted to minimize federal powers and inquiry, while the North did not. Thus, the northern Congressmen pressed for (and largely succeeded in getting) an expanded census, over the objection of their southern counterparts, who feared that information collected on their constituencies would be used to discredit slavery.

This sectional conflict reached its fullest expression in the debate over Schedule No. 2, the slave schedule. If the Census Board had had its way, present-day social scientists would know considerably more about the African American population of the mid-nineteenth century. The Census Board proposed that Schedule No. 2 include inquiries on the names of slave owners; the number of slaves; the age, sex, and color of each slave; whether the slave was a fugitive from the state; the number of manumitted slaves; and the number of deaf and dumb, blind, insane, or idiotic slaves. The proposed schedule also included queries regarding the names of individual slaves; the birthplace of each slave; the number of children born to each female slave and the number of those known to be alive or dead; and the "degree of removal from pure blood" of each slave. The discussion of these last four questions degenerated into a heated fight that divided the floor along sectional lines.[7] Unfortunately for historical researchers, southern concerns won this particular battle. The slave schedule that then gained congressional approval included only: names of slave owners; number of slaves; age; sex; color; fugitives from the state; number manumitted; and whether deaf and dumb, blind, insane, or idiotic.

The 1860 schedule ignited little recorded debate in Congress or the broader statistical community, and the 1850 census legislation and schedules applied to the 1860 count almost without exception.[8] But the 1870 census was contested. The players here included an obstinate Forty-first Congress, an organized statistically oriented elite, and a Special Committee of the House of Representatives, headed by James A. Garfield and including "expert" advisors such as Francis A. Walker and Dr. Edward Jarvis.

Garfield et al. proposed to amend existing population schedule questions on race, school attendance, literacy, and the disabled. The committee also pitched the adoption of new questions: on the relationship of each person to the head of the family; marital status; language spoken; religion professed; whether the individual's parents were of foreign birth; and construction material and value of each dwelling.[9] These proposed new inquiries did not derive from changing social conditions in the United States per se. Rather, they followed the recommendations of an international community of statisticians comprised of "the pro-

FIGURE 1
1840 Census of Population Schedule

Name of county.	Name of ward, town, township, parish, precinct, hundred, or district.	FREE WHITE PERSONS.																						FREE COLORED PERSONS.										SLAVES.										Total.
		Males.										Females.												Males.										Females.										
		Under 5.	5 and under 10.	10 and under 15.	15 and under 20.	20 and under 30.	30 and under 40.	40 and under 50.	50 and under 60.	60 and under 70.	70 and under 80.	80 and under 90.	90 and under 100.	100 and upward.	Under 5.	5 and under 10.	10 and under 15.	15 and under 20.	20 and under 30.	30 and under 40.	40 and under 50.	50 and under 60.	60 and under 70.	70 and under 80.	80 and under 90.	90 and under 100.	100 and upward.	Under 10.	10 and under 24.	24 and under 36.	36 and under 55.	55 and under 100.	100 and upward.	Under 10.	10 and under 24.	24 and under 36.	36 and under 55.	55 and under 100.	100 and upward.	Under 10.	10 and under 24.	24 and under 36.	36 and under 55.	55 and under 100.	100 and upward.	

Source: Ninth Census Report, 41:2, House Report 3, 1869 (pp.41-42).

FIGURE 2
1850 Census of Population Schedule

Dwelling-houses numbered in the order of visitation.	Families numbered in the order of visitation.	The name of every person whose usual place of abode on the first day of June, 1850, was in this family.	DESCRIPTION.			Profession, occupation, or trade of each person, male and female, over 15 years of age.	VALUE OF ESTATE OWNED.*		Place of birth, naming the state, territory, or country.	Married within the year.	Attended school within the year.	Persons over 20 years of age who cannot read and write.	Whether deaf and dumb, blind, insane, idiotic, pauper, or convict.
			Age.	Sex.	Color. { White, black, or mulatto.		Value of real estate.	Value of personal estate.					
1	2	3	4	5	6	7	8	9	10	11	12	13	14

* In 1850 the schedule had but one column for "value of real estate owned." The ninth column was added in 1860, which extended the number from 13 to 14.

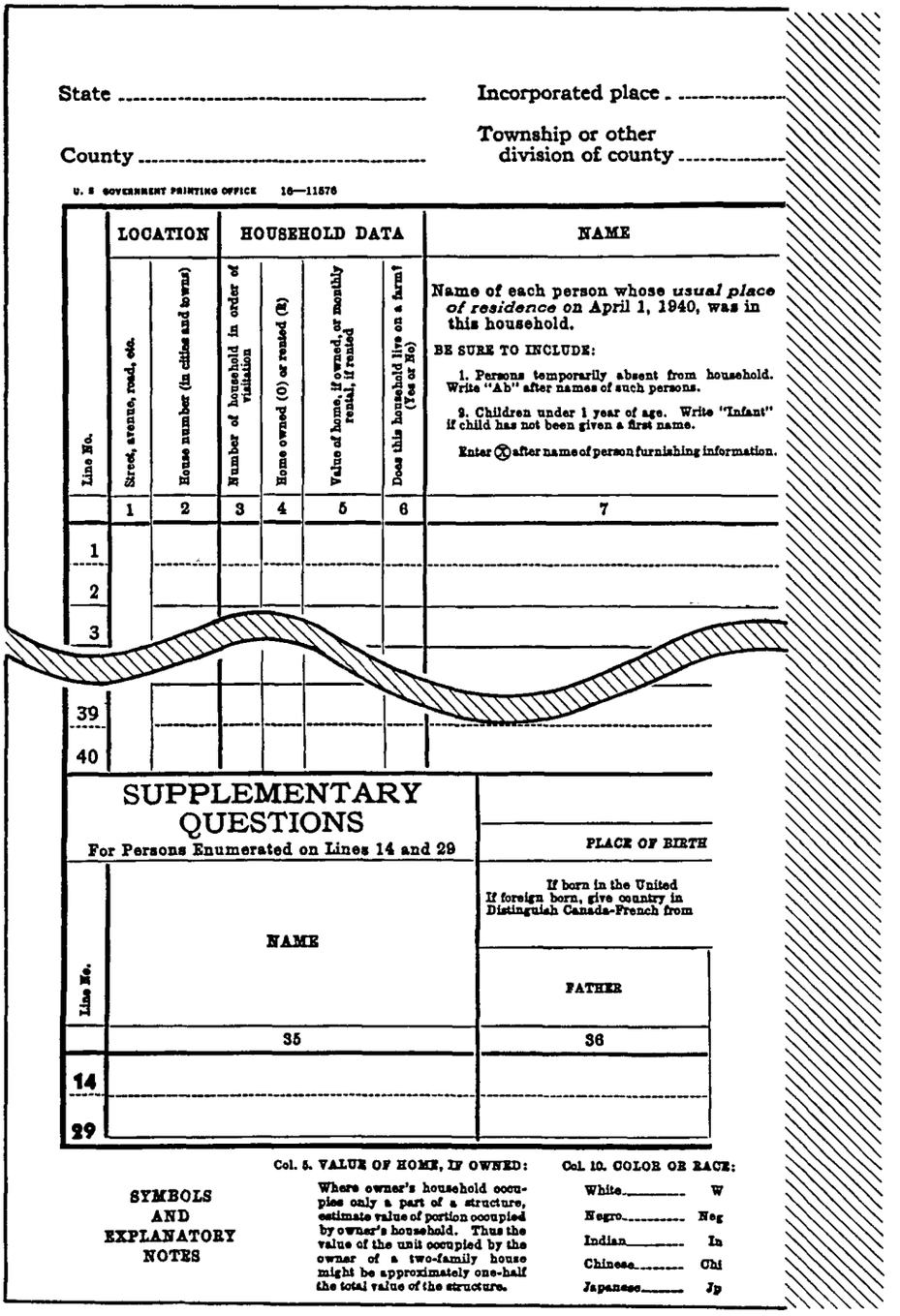
Source: Ninth Census Report, 41:2, House Report 3, 1869 (pp.41-42).

foundest scholars of Europe and America." Via an international congress, these nascent social scientists had developed lists of "indispensable" and "recommended" questions and had encouraged lobbying for their inclusion in national censuses. By contrast, the proposed addition of a "constitutional relations" question addressed a uniquely American political problem. Because Congress viewed the census as the only means of determining apportionment based on the Fourteenth Amendment, a question regarding the abridgment of citizenship rights was a suggested addition to the population schedule.[10]

Of all these proposed changes, only the last gained congressional approval for the 1870 count. Congress could not accept that the 1850 census act, "intended as an enduring provision," had already "become obsolete before the third occasion" arrived for its use.[11] An unusually large dose of skepticism and mistrust, even for a census bill, plagued the Senate during its debate on the 1870 census. The argument for continuity, espoused by Senator Conkling of New York, proved persuasive: "... why keep inquiries after the same things? Because the same kind of statistics time after time admit of certain and easy comparisons. Changes of growth or decline, and the lessons changes teach, constitute the chief object with the students of statistics. Changes can be defined and measured only by comparisons, and he who destroys uniformity among tables which belong to a series, is an enemy of science and of knowledge."[12] The lengthy and cogently argued report of the Special Committee, and the backing of the national and international statistical communities, could not overcome apathy, stubbornness, and the personal vendettas of some congressmen. The census of 1870 was taken under the 1850 census act, with minor amendments.[13]

Like the census of 1850, the 1880 enumeration marks a fundamental departure from its predecessors. By 1880, two factors favored sweeping changes like those that had been proposed by Garfield and Walker for 1870. First, Congress was willing to listen to the concerns of the national and in-

FIGURE 3
1940 Census of Population Schedule



State _____ Incorporated place _____

County _____ Township or other division of county _____

U. S. GOVERNMENT PRINTING OFFICE 16—11576

Line No.	LOCATION: Street, avenue, road, etc.	LOCATION: House number (in cities and towns)	HOUSEHOLD DATA: Number of household in order of visitation	HOUSEHOLD DATA: Home owned (O) or rented (R)	HOUSEHOLD DATA: Value of home, if owned, or monthly rental, if rented	HOUSEHOLD DATA: Does this household live on a farm? (Yes or No)	NAME: Name of each person whose *usual place of residence* on April 1, 1940, was in this household. BE SURE TO INCLUDE: 1. Persons temporarily absent from household. Write "Ab" after names of such persons. 2. Children under 1 year of age. Write "Infant" if child has not been given a first name. Enter ⊗ after name of person furnishing information.
	1	2	3	4	5	6	7
1							
2							
3							
39							
40							

SUPPLEMENTARY QUESTIONS

For Persons Enumerated on Lines 14 and 29

Line No.	NAME	PLACE OF BIRTH: If born in the United If foreign born, give country in Distinguish Canada-French from: FATHER
	35	36
14		
29		

SYMBOLS AND EXPLANATORY NOTES

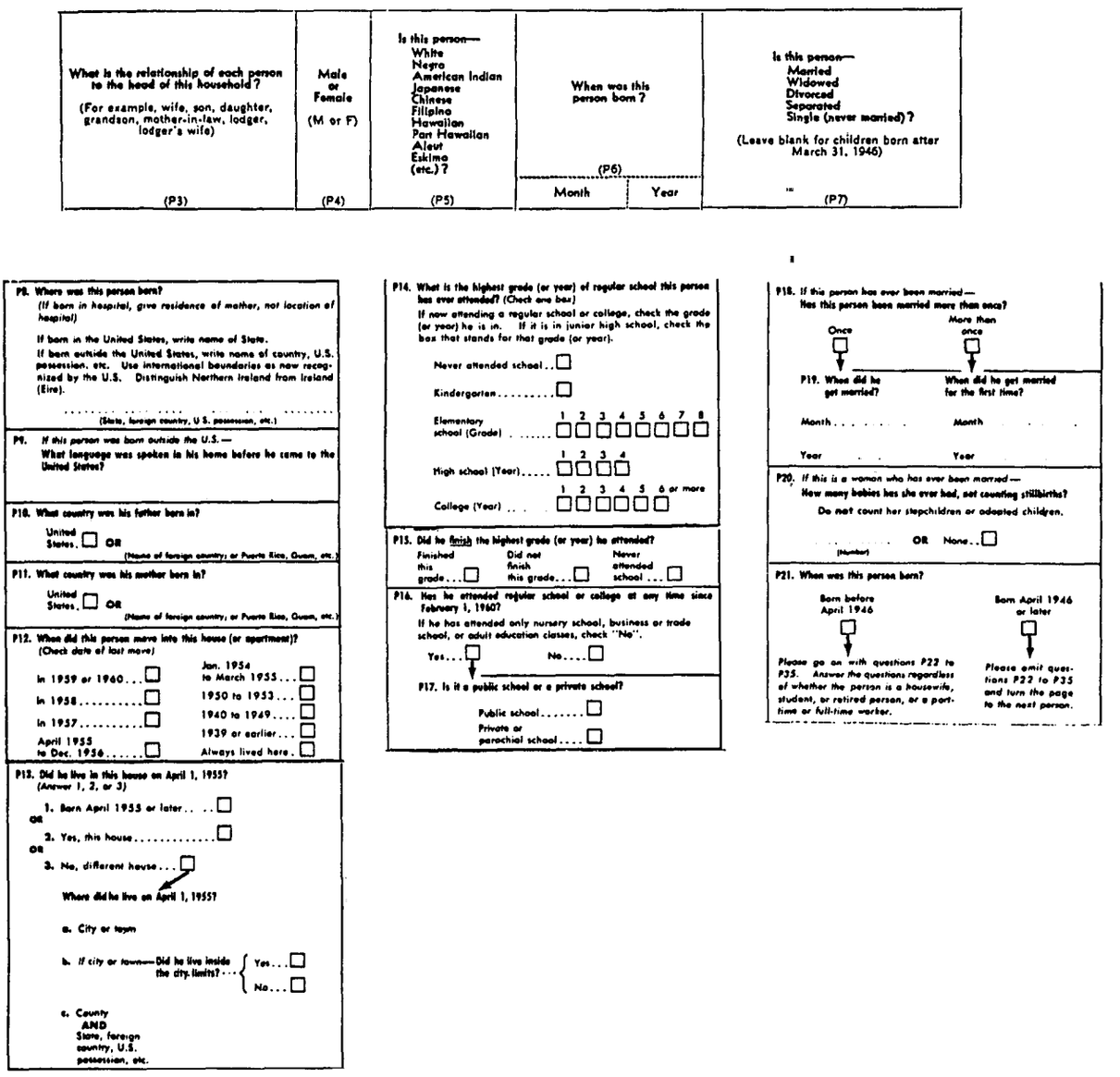
Col. 5. VALUE OF HOME, IF OWNED: Where owner's household occupies only a part of a structure, estimate value of portion occupied by owner's household. Thus the value of the unit occupied by the owner of a two-family house might be approximately one-half the total value of the structure.

Col. 10. COLOR OR RACE:

White	W
Negro	Neg
Indian	In
Chinese	Chi
Japanese	Jp

Source: U.S. Bureau of the Census (1989), *200 Years of U.S. Census Taking: Population and Housing Questions, 1790-1990* (Washington: GPO).

FIGURE 4
1960 Census of Population Schedule



Source: U.S. Bureau of the Census (1989), *200 Years of U.S. Census Taking: Population and Housing Questions, 1790-1990* (Washington: GPO).

ternational statistical communities. Second, Walker, retaining his job as the Superintendent of the Census, was in a better position to argue for a massive overhaul of census taking.

Superintendent Walker personally influenced change in the 1880 population schedule. The questions on the value of real and personal property owned are a case in point. In his report to the Secretary of the Interior, Walker had inveighed against the queries on the value of property owned.[14] He argued that the questions caused "much irritation and annoyance" and the responses were unreliable. It was generally the woman, her children, or the hired help who answered the enumerator's questions. Such persons, Walker insisted, could not be expected to know the answers to inquiries about wealth and property. Apparently Congress agreed, amending the 1879 census act to omit the questions on real and personal property.[15] The population questions

added in 1880 concerned relationship to head and marital status (the two "indispensable" queries urged for 1870) and the number of months unemployed during the census year.[16]

Eighteen-eighty marks the beginning of increased lobbying for specific questions to be added to, refined, or dropped from the decennial census by academics, congressmen, reformers, businesses, and interest groups.[17] This is an important shift in the factors influencing which questions do or do not appear on the population schedule. By the end of the nineteenth century, as the census attracted broader national interest, its original constitutional purpose as a determinant of apportionment was eclipsed by its production of social statistics. The census now had a public, and this public demanded more and better questions.[18]

The growing strength of specific interest groups in formulating census questions is demonstrated by the supporters of war veterans. A question on veteran status, last seen on the 1840 census, was updated for an appearance on the 1890 schedule.[19] An 1888 memorial urged this addition, to serve veterans and "officials" concerned with veterans' affairs and to gather "material for computing the cost of a service pension."[20] Veterans' affairs supporters would continue to hound the Census Bureau for statistics on veterans at each census throughout the twentieth century (see table 1).

Despite the Bureau's growing authority over schedule content and the increasing pressure of interest groups, Congress still maintained the power to impose questions on the population schedule that it deemed important. A special act passed by Congress in 1890 required the census superintendent to collect information on the status of the farm population. The terms of this act mandated adding six new questions to the population schedule.[21]

The greatest innovation of the 1890 population schedule was the introduction of individual family schedules.[22] The population schedules for earlier censuses were large pages accommodating entries for forty to fifty individuals per sheet; the smaller 1890 form allotted a separate schedule to each family enumerated. The adoption of the family schedule in 1890 permitted experimentation with its use as a *prior* schedule, which had been used successfully in Britain since 1851. Such schedules derived their name from the practice of distributing the schedules to all known households prior to the official date of enumeration. The enumerator returned on enumeration day or soon thereafter to retrieve the completed schedules and answer any questions.[23] Despite initial praise for this innovation, the use of separate forms for each household was cumbersome, and the 1900 population schedule returned to the familiar format of lines for fifty individuals per sheet.[24]

By 1899, the climate of opinion at the Census Office and in Congress surrounding the decennial population count was less than enthusiastic. There was "general appreciation of the fact" that the decennial censuses of 1880 and 1890 had gone too far in their zeal to collect social statistics, sacrificing data quality on the altar of data quantity. The mood of the Census Office emphasized "getting back to the basics" for the 1900 census. Understandably, then, the 1900 population schedule dropped eight inquiries, refined seven, and added only two (see table 2). The 1910 schedule was similarly conservative in the number of additions to the population schedule.

For the new date of birth question, the enumerators of 1900 were to record the month and year of every individual's birth, to increase the accuracy of statistics on age. Supporters of this question argued that "many a person who can tell the month and year of his birth will be careless or forgetful in stating the years of his age."[25] The inquiry failed to satisfy these hopes, however. After the tabulation of the age data had begun, Census Bureau Statistician Joseph A. Hill became suspicious of the returns after carefully examining original schedules from different sections of the country. He thereby determined that the date of birth query was often "computed by the enumerator and entered at the close of the day's fieldwork" and "by a large proportion of the enumerators the reply to the question of age was assumed to be correct, while the year of birth was made to correspond with it."[26] Even worse, Hill discovered that in many cases the enumerator's computation of the date of birth was incorrect. Concluding that the addition of the date of birth question in 1900 did not increase the accuracy of that year's age statistics, Hill argued persuasively that it be dropped from the 1910 population schedule.[27] The second new question—number of years married—replaced the "whether married during the census year" query, which had appeared in every decennial census of population since 1850.

The most striking change for the 1900 population schedule was the omission of questions relating to health and dependency. Questions on health status had been part of the decennial census of population since 1830. By 1900, however, gone were the inquiries on acute or chronic disease; whether crippled, maimed, or deformed; as well as the queries on insanity and idiocy. Not even the old standby "deaf and dumb" question appeared on the 1900 schedule. Also dropped from the 1900 census was the inquiry regarding "whether a prisoner, convict, homeless child, or pauper."

The testimony of Director of Census Simon N.D. North explains these omissions: "Every enumerator was required to ask those . . . questions in every family that he visited, and there was a terrible uproar about it." North continued, "The difficulties encountered by the enumerators were so serious that practically the questions were withdrawn before the census was completed. The unfortunate experience made it seem desirable to those concerned with the Twelfth Census to omit questions of this character from the schedule."[28] Another rationale was limiting the total number of questions. As North wrote to one supporter of the deaf, blind, and dumb inquiries, "The difficulties and complications attending the decennial enumeration of the population

for the United States are increasing so rapidly that it has become imperative to simplify and reduce the range of the inquiries."[29] Congress heeded North's pleas to drop the health queries from the 1900 schedule.[30] Nonetheless, North was forced to revive the questions on the deaf, dumb, and blind for the 1910 population schedule, against his best judgment.[31]

Just before the 1920 census, a new player formally entered the circle of influence over census inquiries—the Joint Census Advisory Committee. This body formalized the long-term cooperative relationship between the Census Bureau and the American Economic Association and the American Statistical Association. The cooperation between the Advisory Committee and the administration of the Census Bureau forms the core of the story of the changing population schedules for the Fourteenth through Sixteenth decennial censuses (1920–40).

While the views of the Joint Census Advisory Committee were particularly influential, this group was one of many trying to shape census schedule content. If the decennial census of 1880 marked the beginning of increased lobbying for specific questions to be added to, refined, or dropped from the population schedule by academics, congressmen, reformers, businesses, and interest groups of all kinds, then the 1920 census demarcates an increasing proliferation of such requests. As had been true from the mid-nineteenth century, general disagreement centered around the scope of the enumeration: those who wanted to limit the number of inquiries versus those who wanted to know more. The Joint Census Advisory Committee and the Bureau were clearly on the side of census simplification.[32]

Memos and letters between the Director of the Census, his Chief Statisticians,[33] and the Chairman of the Joint Census Advisory Committee agonized over the pressures to include new questions and refine old ones on the population schedule. The weariness of the Chairman of the Joint Census Advisory Committee over the deluge of requests is evident in his note to the Director of the Census in 1928: "To add to your joys I am sending you herewith some correspondence from the American Statistical Association with some more of these demands for additional inquiries."[34] The Joint Census Advisory Committee took on the job of sifting through the multitude of suggestions for questions and passing along the recommendations to the Census Bureau. By 1939, the Bureau was bemoaning the thousands of requests it received on adding, refining, or reinstating questions.

Records of a committee organized by the Bureau to wrestle with schedule issues reveal the Census Bureau's internal decision-making process regarding the 1920 population schedule. The Committee on Legislation for the Fourteenth Census, composed of the Director of the Census, a secretary, and the six Chief Statisticians, provided a forum for dialogue about census schedule content and was in constant communication with the Joint Census Advisory Committee. After the Committee on Legislation concluded its meetings, its suggestions were submitted to Congress.

In 1920, the Census Bureau administration clearly placed a high priority on reducing the number of questions on the population schedule, even at the cost of eliminating inquiries that had persisted across many decades. At the fifth meeting of the Committee on Legislation in August 1917, eleven questions from the 1910 population schedule were on the chopping block.[35] It took only an hour and forty minutes for the committee to agree unanimously to eliminate all eleven questions: relationship to head of family; number of years of present marriage; mother of how many children, number born and number now living; if an employee, whether out of work on 15 April 1910 and number of weeks out of work during year 1909; farm or house; number of farm schedule; whether a survivor of the Union or Confederate Army or Navy; whether blind (both eyes); and whether deaf and dumb. The query on illiteracy was saved from elimination only by the intercession of the Director and Chief Statistician for Population, who supported maintaining a question allowing "harmonious" comparisons from decade to decade.[36]

The minutes of the 1920 Committee on Legislation reveal just how important the views of individuals were in deciding schedule content. For example, Chief Statistician of Population William C. Hunt was largely responsible for the appearance of questions on illiteracy, mother tongue, and relationship to the household head on the 1920 population schedule. Two months after the Committee on Legislation voted to remove the relationship to head question, Hunt urged its reinstatement because "this item was very useful as a check on the accuracy of other data on the schedules." He convinced all his colleagues except the Bureau geographer, who "opposed its restoration on the grounds that it was of little or no value."[37] It is clear that the strategic placement of key individuals—not just broad "social forces" acting on the population schedule—made a big difference in outcomes.

The pressures on the Census Bureau to include new questions and omit or refine others on the 1930 and 1940 population schedules were even greater than in previous decades. Outside pressure to increase the scope of enumeration came indirectly from Section 4 of the Census Act of 18 June 1929, which authorized the Director of the Census to select the particular items to be included in the population schedule, subject to approval by the Secretary of Commerce.[38] The Census Bureau and the Joint Census Advisory Committee agreed that the population schedule should be moved "in the direction of simplification." According to the Bureau and the Advisory Committee, a leaner schedule would not only simplify enumeration and tabulation, but it would also increase the accuracy of the count.[39] The interested public, however, continued to protest as long as "their" question remained off the population schedule. It would be the experience of the Census Bureau that question boosters, if persistent and loud

enough, clearly posed a threat to the Director's authority over what would appear on the population schedule.[40]

Beginning with the 1930 census, and continuing to the present, the Census Bureau has relied on formal procedures to elicit, evaluate, and integrate the views of the public on population schedule content. "Everyone who had even a remote interest in the matter" was invited to Washington to attend a Bureau-sponsored conference.[41] In this structured atmosphere, public opinion was gleaned through the suggestions and criticisms of the representatives. The conference was intended "to draw the poison, and to get expressions of opinion from those in attendance," so that "for all time" if anyone cried foul over the schedule content, the Bureau could respond, "Why didn't you attend . . . and express yourself on the question?"[42]

The Bureau had obviously used the conferences as a form of appeasement to its public; it wanted to appear receptive and cooperative to interested organizations and individuals and to give them their "day in court." Believing this psychology of public openness "spelled the difference between a successful census and one that was a failure,"[43] the Bureau encouraged interested parties to voice their concerns and physically work on the schedule in cooperation with the Bureau. Ideally, the final product would be met with the feeling that both sides had compromised and contributed to the end result. Another reason the Bureau used a population conference was political. Past experience taught Bureau administration that if interested individuals and organizations did not feel they had been given a fair hearing, they would make suggestions directly to the Secretary of Commerce and even the White House.[44] Even though the Director of the Census had the final authority over the population schedule, the Bureau wanted to insure itself against this kind of pressure.

The inclusion of unemployment questions in 1930 illustrates a growing tension between the Census Bureau and Congress in the twentieth century. The Census Bureau understood its work in terms of a continuum or as a statistical whole; certain questions were so important they were asked in every census, and new questions usually reflected population concerns that had been articulated for years before their inclusion. In the twentieth century, however, Congress increasingly viewed the census as a tool for understanding immediate concerns, such as joblessness during the depression.

The story of who and what determined U.S. population schedule content in the last half of the twentieth century has continued to revolve around this tension. The advent of expanded federal grants-in-aid, civil rights laws, and reapportionment decisions burdened the Census Bureau in new ways.[45] Increasingly, questions were included in the census not because the Bureau found them inherently interesting or important to the statistical and academic communities, but because they were mandated by federal legislation. Beginning in the 1960s and continuing to the present, an enormous amount of federal funding has been allocated to communities based on characteristics enumerated by the census (e.g., number of rooms, year structure built).[46]

The 1990 decennial census is the first census in nearly a century to omit questions on the age, duration, or timing of marriage. In a period when marital patterns are changing more than ever before, this is a curious omission. Further research is necessary to uncover the persons affecting this change, but it is clear that these important demographic questions lacked a strategically placed individual or group championing their inclusion on the population schedule.

NOTES

1. All the issues covered in this article are treated with greater detail by Diana L. Magnuson (forthcoming dissertation). Portions of this article appeared in King and Magnuson (1993).

2. Anderson (1988).

3. Cohen (1982); Anderson (1988).

4. The Census Board comprised the Secretary of State, the Attorney-General, and the Postmaster-General. The Board was required by this act to "prepare and cause to be printed such forms and schedules as may be necessary for the full enumeration of the inhabitants of the United States; and also proper forms and schedules for the collecting in statistical tables, under proper heads, such information as to mines, agriculture, commerce, manufactures, education, and other topics as will exhibit a full view of the pursuits, industry, education, and resources of the country; it being provided that the number of said inquiries, exclusive of the enumeration, shall not exceed one hundred, and that the expense in preparing and printing said forms and schedules shall not exceed $10,000." 3 Mar. 1849, *Statutes at Large* 9.

5. Free population, slave population, mortality, agriculture, industry, and social statistics.

6. *Congressional Globe* 30 (3): 627–29.

7. "Seventh Census," *Congressional Globe* 31(1): 673–77.

8. Wright and Hunt (1900,50–52); Holt (1929,18).

9. *Ninth Census Report*, 41:2, House Report 3, 18 Jan. 1870.

10. The "constitutional relations" question was not suggested by the Select Committee on the Ninth Census.

11. "Ninth Census," *Congressional Globe* 41(2): 1079.

12. Ibid.

13. 23 May 1850, *Statutes at Large* 9; 6 May 1870, *Statutes at Large* 16.

14. *Annual Report of Superintendent of Census* 45:3, *House Exec. Doc. No. 1* (17 Jan. 1878): 849.

15. *Compendium* (1854, 21–23); "Ninth Census," *Congressional Globe* 41(2): 1106.

16. The other additional questions urged by Garfield et al. for 1870 were not added. Eighteen-eighty is more remarkable for its change in enumeration procedure than for its change in schedule content.

17. See, for example, "Memorial of Mary F. Eastman, Henrietta L.T. Wolcott, and others, Officers of the Association for the Advancement of Women, praying that the Tenth Census may contain a just enumeration of women as laborers and producers," 45:2, *Senate Misc. Doc.* 84; *Cong. Rec.* 51 (1): 2978–83; 3096–97; 3186. See also Anderson (1988).

18. Anderson (1988).

19. "Whether a soldier, sailor, or marine during the civil war (U.S. or Confederate), or widow of such person."

20. "Memorial of Henry Hall, of New York, urging legislation that shall incorporate in the next census provisions for taking a complete enumeration of the surviving veterans of the war of the rebellion, including names, age, residence, length of service, and the commands under which they served," 50:2, *Senate Misc. Doc. No. 26*, 2 Jan. 1889.

21. The six new questions required by the act: the number of persons who live on and cultivate their own farms; the number who live in their own homes; the number who hire their farms and homes; the number of farms and homes which are under mortgage; the amount of mortgage debt; and the value of the property mortgaged. The schedule was further required to determine "whether such farms and homes have been mortgaged for the

whole or part of the purchase money for the same, or for other purposes, and the rates of interest paid upon mortgage loans." 22 Feb. 1890, *Statutes at Large* 26; "Resolution of Inquiry Relating to the Census," 51:1, *House Misc. Doc. No. 46*, 6 Jan. 1890.

22. The use of prior schedules in the United States had been tossed around for quite some time. Superintendent of Census J.D.B. DeBow had noted the general success of such schedules in Britain. In the Forty-first Session of Congress the idea of distributing prior schedules was defeated along with the rest of Garfield's proposed census law. *Compendium* (1854, 21–23); "Ninth Census," *Congressional Globe* 41 (2): 1106.

23. Ibid.

24. *Annual Report of Superintendent of Census* 51:1, 30 June 1889, *House Exec. Doc. No. 1*: 15–16. *Population* (1902, ccxxi).

25. National Archives, Record Group 29, Entry 200, Box 150, File: P-7 (hereinafter NA RG 29, 200/150: P-7).

26. Much to his disgust, Hill found that in some cases the entire column of year of birth on a schedule was in a different ink than that used for the remainder of the schedule. NA RG 29, 200/150: P-7.

27. Ibid.

28. *Cong. Rec.* 60 (2): 1153.

29. Ibid.

30. Provision was made for some disabled to be enumerated through a special schedule, however. *Cong. Rec.* 60 (2): 1153.

31. North received numerous letters requesting the reinstatement of queries regarding the blind and the deaf and dumb as well as other disabled classes. *Cong. Rec.* 60 (2): 1153.

32. NA RG 29, 148/72; Advisory Committee, Dec. 1928.

33. The Chief Statisticians included a Chief Statistician of population; agriculture, cotton, and tobacco; manufactures; statistics of cities; vital statistics; and revision and results.

34. NA RG 29, 148/72; Advisory Committee, May 1928.

35. NA RG 29, 205/2.

36. NA RG 29, 149/2117.

37. NA RG 29, 205/2.

38. Previous census acts stipulated that Congress list the questions to be included in the decennial census of population. After 1930, Congress merely specified the areas to be investigated.

39. NA RG 29, 148/72; Advisory Committee, Dec. 1928.

40. The Bureau could blame itself in part for the enormous number of letters it received suggesting questions for the decennial population schedule. During 1928 and 1929, the Census Bureau invited the public to contribute input for the 1930 census questions and employed the services of unofficial advisory committees outside the Bureau to help sort out the enormous volume of responses its solicitation generated. Oliver McKee Jr. (1930).

41. NA RG 29, 148/74; Advisory Committee, Feb. 1939.

42. Ibid.

43. NA RG 29, 148/74; Advisory Committee, Sept. 1938, Feb. 1939.

44. NA RG 29, 148/74; Advisory Committee, Feb. 1939.

45. Anderson (1988, 213).

46. *Population Today* (1993, 10).

PART 1. Historical Comparability of the U.S. Census

Comparability of the Public Use Microdata Samples: Enumeration Procedures

Diana L. Magnuson and Miriam L. King

An important comparison of the Public Use Microdata Samples (PUMS) is that of the enumeration procedures used to collect the data in the first place. Spanning 150 years, the Integrated Public Use Microdata Series (IPUMS) covers eleven censuses and represents at least as many different methods of enumeration. This article concentrates on three issues of procedural comparability: the quality of local administrators, the selection of enumerators, and the training and oversight provided by the national census office.[1]

Since 1950, the U.S. Bureau of the Census has published its own official procedural histories (U.S. Bureau of the Census 1955, 1966, 1976, 1986–1989, 1993). For the 1940 census, users of the public use sample have relied on the procedural history created in conjunction with the 1940 PUMS (Jenkins 1985). With the exception of 1880 (King and Magnuson 1993), no procedural histories exist for the censuses before 1940, forcing social scientists to refer to diverse secondary source material (e.g., U.S. Census Office 1882; Walker 1888; American Economic Association 1899; Wright and Hunt 1900; Holt 1929; Eckler 1972; Anderson 1988). This article focuses primarily on the censuses of 1850–1940, giving only cursory attention to the censuses of 1950–1990, for which procedural histories already exist.

As was the case for earlier censuses, the 1850–1870 decennial censuses of population were conducted under the supervision of the marshals of the judicial districts of the United States. Marshals had the power to appoint "as many assistants within their respective districts as to them shall appear necessary." The characteristics of the assistants were to be those of "assiduous industry, active intelligence, pure integrity, great facility and accuracy of computation" and "an intimate knowledge of the division allotted to them."[2]

It seems that the assistant marshals often fell short of these desiderata. In 1854, Superintendent of Census J.D.B. DeBow complained about the kinds of men who were likely to gain appointments as assistant marshals. These men, he argued, were not selected "for their especial fitness," but rather because they were "willing to undertake it."[3] Former Superintendent of Census Kennedy affixed blame for this state of affairs on the low rate of compensation for services rendered. A second problem was that positions were commonly awarded as rewards for services to a political party, rather than according to the candidates' qualifications as canvassers.

Another problem was the lack of training and oversight to guide assistant marshals. While the establishment of the Census Board and the creation of the Department of the Interior in 1850 dramatically improved some aspects of census administration, the actual procedure for canvassing the population remained virtually unchanged from the first census. The 1850 census superintendent did improve upon previous practice by sending to marshals circulars "explaining and defining each inquiry, in addition to general instructions and printed schedules."[4] Marshals were also required, through personal correspondence with the office of the Secretary of the Interior, to explain any discrepancies in their returns. The paucity of procedural innovations in the 1850–1870 censuses did not preclude prodding for major changes by the Superintendent of the Census, however. By

1870, the Census Bureau administration and other interested parties were aggressively advocating change in the methods of enumeration that had been used for eighty years.

Kennedy's complaints were echoed with greater tenor in 1869 by the Special Committee of the House of Representatives on the Ninth Census, headed by future President James A. Garfield and supported by Francis A. Walker.[5] While the efforts of the Special Committee on the Ninth Census ultimately "came to naught" in 1870, they reached fruition with the 1880 census. By 1879, two factors favored those advocating procedural reform. First, in the decade between the Ninth and Tenth censuses, Walker gained credibility as the country's leading demographic expert and fine-tuned his arguments favoring revamping enumeration procedures. In addition, congressional sentiment toward procedural reform had caught up with the enthusiasm and concern of Walker and those represented by the Special Committee. When Garfield again spoke on behalf of the reforms he had proposed in 1869, Congress agreed that the machinery of the census was in need of "changes so great as to amount to revolution."[6]

The 1880 census marked a turning point in local administration, enumerator selection, and oversight by the Census Bureau.[7] If the 1850 census qualifies as "the first modern census" on the basis of its population schedule, the 1880 census merits the same title in terms of its innovations in enumeration procedures.

At the top of Walker's list of procedural reforms for the 1880 census was the elimination of marshals as overseers of the United States census. Walker pointed out that marshals, who gained their posts through political patronage, had as their primary duty the apprehension and control of criminals, not the enumeration of the American populace. The burden of census work was too great to be performed adequately by officials who were already "crowded to the limits of their time and strength by prior official duties."[8] A second advantage of replacing marshals under Walker's scheme was that the new officials would be under the direct control of the Census Office and the Department of the Interior, while marshals were answerable to the Department of Justice. Finally, these new officials could be selected because they had qualifications suited to the particular requirements of administering the census.[9] Upon their shoulders would rest the responsibility to "select, appoint, commission, instruct, supervise, and finally correct the work of" enumerators.[10]

The argument that U.S. marshals had "neither the facilities nor the necessary qualifications to make a complete and accurate enumeration" finally got through to the Forty-sixth Congress.[11] Under the 1879 census act, Congress replaced marshals with officials known as supervisors. The Secretary of the Interior was given the responsibility of designating 150 supervisors, representing the states and territories, to be approved by the President "by and with the advice and consent of the Senate."[12] Walker's point that those who supervised census taking should not be burdened with other duties was sound. After the 1880 count was completed, supervisors adamantly stressed the time-consuming and demanding nature of their post.[13] Given other competing duties, marshals could hardly have devoted the time and care to overseeing the census that the supervisors claimed to have done in 1880.

Whether the administrators of census taking were under the control of the Department of Justice (as U.S. marshals) or the Department of the Interior (as census supervisors) may seem a minor detail. Nonetheless, the Superintendent argued convincingly that the power to appoint and remove supervisors was critical to the quality of the enumeration.[14] The absence of this pressure had, in Walker's view, vitiated the quality of the 1870 enumeration. The Census Office then had no power to veto the marshals' plans for dividing up territory, and several marshals had insisted "against the advice of the Census Office . . . on assigning to assistant marshals districts which could not possibly be canvassed in compliance with law in the prescribed time, the result being either the undue protracting of the enumeration, or else the illegal letting out of the work to unauthorized parties."[15] Similarly, lacking the power to veto enumerator appointments, the Census Office could only express "its entire disapprobation" at the appointment of southern enumerators in 1870 "whose appointment was disgraceful to the government and detrimental to the service," leading to "mischievous and even scandalous results."[16]

One of the most important tasks of the census supervisors was the selection of enumerators.[17] The census law and Census Superintendent set general guidelines for these officials to follow in choosing canvassers. Nonetheless, press coverage of local census taking indicates that supervisors and communities varied in their implementation of these guidelines.[18] Across the country, however, similar factors encouraged the appointment of qualified enumerators. Supervisors forwarded their lists of names to Washington for review, and the Census Office could deny (and sometimes did) confirmation to incompetents and "political hotheads."[19] Supervisors swore to discharge their duties in accordance with the census law (which demanded enumerators be chosen on the basis of competence) and were liable to two years' imprisonment for any dereliction of duty.[20] Finally, the most effective safeguard on the quality of enumerator appointments was a supervisor's awareness that hiring the unfit would reduce his community's count and incur the ire of the citizenry.[21]

The "changes so great as to amount to revolution" in enumeration procedures (that Garfield had spoken of in his speech to Congress) were manifested in the Census Bureau's concerted effort to train and oversee census supervisors and enumerators in 1880. Walker was painfully aware that no matter how carefully enumerators might be selected, the quality of their returns was shaped by the mechanisms in place to "instruct, supervise, and finally cor-

rect" their work.[22] The responsibility for defining and carrying out these tasks for the Tenth Census was split between the Census Office in Washington, D.C., and the district supervisors.[23] Although these procedural changes greatly increased the preliminary enumeration work of the Census Office, the Superintendent's opinion was that "no one should be deemed fit for such a charge who did not rejoice in the added labor and care, in view of the manifold advantages to be obtained."[24]

In early May of 1880, supervisors received from Washington advance samples of the schedules; in late May, the actual blanks arrived. For their guidance, enumerators also received a letter of instruction that spelled out pay rates in their district; commissions and oaths; a pamphlet containing "clear and minute" instructions on how to conduct themselves and fill out the schedules; a completed sample schedule; and, in some districts at least, copies of the census law.[25] While the enumeration was under way, supervisors were required to be in constant contact with enumerators, communicating "the necessary instructions and directions relating to their duties, and to the methods of conducting the census," as well as advising and counseling enumerators "in person and by letter, as freely and fully as may be required to secure the purposes" of the law.[26] The census act further directed supervisors to "examine and scrutinize" enumerator returns.[27]

Enumeration procedures put in place for the 1880 census continued to serve as a model for succeeding U.S. censuses well into the twentieth century. In other domains, the Census Bureau implemented major changes, such as introducing automatic tabulating machinery to process the returns of the 1890 census. But in the areas of local administration, enumerator selection, and oversight of fieldwork by the Washington office, successive enumerations largely refined and elaborated methods introduced in the Tenth Census. For example, the 1880 model of enumerator selection was significantly improved with the innovation of enumerator exams in 1900. Whereas in 1880 and 1890, supervisors made enumerator appointments based upon their own judgment of the character and competency of each enumerator, beginning in 1900 would-be enumerators were required to submit to a written examination.[28] The test schedule was an exact copy of the population schedule mailed to each candidate and "filled out in hypothetical manner" using a sample narrative.[29] Candidates returned to their respective supervisors the completed test schedule, together with certification that they had not received any assistance in filling it out. Despite the new procedures for procuring enumerators, Bureau officials continued to lament the difficulties in obtaining "properly trained men" for census fieldwork. Indeed, census officials suspected that many candidates received substantial assistance in filling out the test schedules.[30]

An important innovation in the oversight of enumerators—the street book—was instituted at the 1900 census.[31] The Census Office used it to aid in the enumeration of larger municipalities and to counter any complaints against the "completeness and the correctness" of the returns. The street book facilitated organization and completeness of the canvass on the part of the enumerator; canvassers used it to record "the number of families and persons found in each house or building," houses or buildings where no persons were found, and the date visited. In other words, enumerators were "required to account for each and every house, building, or place of abode within the limits of [their] enumeration district. . . ."[32] The street book had another important purpose: supervisors used it to verify the completeness of each enumerator's canvass. Director of the Census William R. Merriam was confident that its use was "a most effective agent" in securing "a much more thorough canvass of city districts than has been possible heretofore." After the enumeration was complete, Director Merriam boasted that the use of the street book was directly correlated to the fact that no large cities had requested re-enumeration.[33]

As in 1900, persons seeking the post of supervisor for the 1910 census were required to apply and were theoretically chosen according to their qualifications. Again as in 1900, would-be enumerators were required to submit an application to their respective supervisors and take a competitive examination. Unlike the 1900 examination procedure, however, anyone wishing to be considered for an enumerator's post in 1910 was required to submit to a proctored examination. Prior to the test date, candidates were mailed census of population and census of agriculture schedules, together with enumeration instructions. Then, on a predetermined test date at locations across the country, candidates took a proctored examination.[34] The successful completion of the examination was not necessarily followed by an appointment. Supervisors were not only required to "give due regard to the relative excellence of the test papers in making their selections," but also to consider those qualifications for an enumerator "which can not be determined by a written test."[35] Those qualifications included age, sex, character, habits, and "standing in the community." While supervisors were "expressly instructed" not to allow partisan politics to play a part in the selection of enumerators, it was generally acknowledged that such practices continued unabashedly. Despite these small gains in the method of obtaining qualified enumerators, the Census Bureau continued to seek "radical" improvements in the process.[36]

The Fourteenth Census Act (1920) departed from the legislation governing the appointment of supervisors from the previous four censuses. Since 1880, supervisors had been appointed by the President, by and with the advice and consent of the Senate. The Fourteenth Census Act, however, stipulated that supervisors were to be appointed by the Secretary of Commerce upon the recommendation of the Director of the Census. Under this arrangement, the Director of the Census had more control over the supervisors. The selection process was also speeded up, giving supervisors time to complete preliminary administrative work prior to census day.[37] For the 1920 and 1930 censuses, the Bureau

issued a press statement outlining the work of census supervisors and calling for interested candidates to request applications from the Director of the Census.[38] In 1920, a "suitable scheme of rating" was devised to rank applicants according to unspecified criteria. From this ranked list, "selections were made" and the list of candidates recommended was sent to the Secretary of Commerce for appointment.[39] In 1930, an extra step was added to this process. Rather than considering every applicant a potential supervisor, the Census Bureau sent a copy of the general instructions for supervisors, along with a cover letter requesting confirmation that the applicant still wanted to be considered for the post of supervisor.[40] The Bureau added this extra step to the appointment process on the assumption that "in many instances persons apply for this position without much idea as to what the duties of the office are, and possibly with the impression that it is something in the nature of a sinecure."[41]

The method of selecting competent enumerators in 1920 generally followed the practice devised in 1900 and 1910.[42] In order to fill all available enumerator positions, supervisors were urged by the Bureau to request "competent persons" or "representative citizens" to apply for enumerators' posts.[43] Those who applied, whether through the urging of a supervisor or of their own accord, were required to take a written exam under the criteria established in 1910.[44] Once supervisors had administered the written exams, they marked, rated, and then mailed them to the Census Bureau where they were examined "as fast as received."[45] Supervisors were then notified promptly as to the number of candidates approved by the Director. From this list, supervisors made their appointments.

Just as 1880 marks a turning point in the administrative machinery of the census at the local and national levels, so too does the census of 1940. The Sixteenth Decennial Census significantly expanded local administration and enumerator training and oversight. After 1940, the fieldwork staff included: a chief of the Field Division; three regional assistants to the chief of the Field Division; 104 area managers assisting each regional assistant; district supervisors under the purview of the area managers; and, finally, the enumerators. The 1940 census also introduced the use of "squad leaders" to coordinate and assist enumerators in cities of fifty thousand or more.[46] These innovations in administrative machinery served as the model to which successive censuses amended minor procedural changes.

The selection and appointment of supervisors and enumerators in 1940 generally followed the procedure outlined in 1930. Unlike the previous three censuses, however, the application method for enumerators in 1940 included an interview. In 1920 and 1930, supervisors interviewed candidates only if the circumstances warranted it (if the candidate could not attend the written examination, for example). In 1940, district supervisors were encouraged to interview applicants "whenever possible" in order to facilitate an atmosphere whereby "applicants could speak freely about their qualifications."[47] After the interview and written examination, each supervisor compiled a list of candidates whom they regarded to be "reasonably expected to qualify" as enumerators. This list became an "eligibility" list from which applicants were "screened."[48] The screening process eliminated "reasonably" qualified applicants who had poor handwriting and those who demonstrated an inability to follow written directions.[49] Once these less-than-desirable applicants were screened from the list, supervisors were instructed to rank applicants according to a hiring preference.[50]

The training and oversight of fieldwork staff expanded significantly with the Sixteenth Census. Supervisors and enumerators were inundated with forms, instructions, memoranda supplements, guidebooks, and manuals. Workshops, training seminars, practice exercises, and constant correspondence were required of fieldwork staff at all levels. Training films, filmstrips, and other audiovisual aids were also used extensively.[51] An expanded hierarchy meant more checks on procedure at all levels of fieldwork.

The 1960 census of population introduced three major changes in enumeration procedure.[52] First, householders filled out the questionnaires. Advance census forms were mailed out to the public ten days to a week before the enumeration date. The forms themselves were a departure from previous censuses. The 1960 schedules were formatted into multiple-choice questions, supposedly to simplify the task of filling out the form. On census day, the enumerator picked up the forms and answered any questions respondents might have regarding the schedule. The Bureau hoped that this scheme of advance census forms, which had previously been used in 1890 (the prior schedule) with relative success, would improve the accuracy of the enumeration. In 1970, the innovation of self-enumeration was taken one step further; householders were instructed to mail the completed census forms back to Washington, thus eliminating the enumerator. After 1960, then, the role of the enumerator shifted to canvassing special populations and follow-up work.

The second change in the 1960 count was the minimal number of questions asked of each individual. One hundred percent of the population received the "short" census form, which asked only five questions. The remaining questions were collected on a 25 percent sample basis on the "long" census form. The 1960 precedent of limiting to under ten the number of population questions asked of the total population, was followed in 1970, 1980, and 1990. In 1970, the Census Bureau used two long forms with somewhat different questions; one sampled 5 percent of the population and the other, 15 percent. The 1980 and 1990 censuses continued to use the long form but sampled different proportions of the population.

Finally, most of the data in 1960 were processed on high-speed electronic equipment. Enumerators transcribed the information from both census forms to specially designed

schedules, which converted the information to machine-readable form by the Film Optical Sensing Device for Input to Computers (FOSDIC), a machine that reads only little circles filled in with No. 2 pencils.[53] Beginning in 1970, householders themselves filled out machine-readable forms.

The census for the year 2000 may usher in a new procedural "revolution." Congress has proposed that the next enumeration collect only those data mandated by a statute. Under this provision, even the long-form schedule completed by a sample of the population could be limited to as few as fifteen questions. In a still more radical departure from past practice, some have called for a shift to "continuous measurement," in which annual surveys would supplant data collection via enumeration once every ten years. It is very likely that the public will be able to answer the Census Bureau's questions in 2000 via telephone or computer.[54]

What generalizations can be made about shifts in enumeration procedures from 1850 to the present? As with changes in the population schedule content, changes in enumeration procedures clustered in specific census years and were followed by periods of fine-tuning an established model. Eighteen-eighty marked a turning point, with the replacement of marshals by supervisors answerable to the Census Office and efforts by the Washington office to provide training, oversight, and correction of fieldwork. By 1940, the complicated administrative hierarchy and flood of directives from the Bureau had standardized enumeration practices across the country. With the shift to a mail census in 1960 and 1970, self-enumeration became the norm; local administration, enumerator selection, and oversight of fieldwork from Washington became less important in determining the completeness and quality of the census.

How much a given enumeration practice or set of practices improved the accuracy and thoroughness of the canvass cannot be specified precisely. Not only enumeration practices but also factors such as the composition of the population and the public's view of the census affected the quality of the count. One summary of quantitative estimates of these various factors is, however, provided by demographic estimates of census undercounts. These figures show similar and fluctuating levels of undercount (between 6.5 and 7.4 percent) from 1880 to 1920, and then a monotonic decline (from 5.3 percent in 1930 to 3.3 percent in 1960 to 1.4 percent in 1980).[55] Such long-term improvement in the quality of the census is consistent with the improvements in enumeration procedures over time.

NOTES

1. All the issues covered in this article are treated with greater detail by Diana L. Magnuson (forthcoming dissertation). Portions of this article appeared in King and Magnuson (1993).

2. 23 May 1850, *Statutes at Large* 9: 428. "Instructions to Marshals—Census of 1820," 20 June 1820, reprinted in Wright and Hunt (1900, 133–37); "Instructions to Marshals—Census of 1830," 15 Apr. 1830, reprinted in Wright and Hunt (1900, 139–41).

3. *Compendium* (1854, 17–18).

4. Holt (1929, 16).

5. *Ninth Census Report*, 41:2, House Report 3, 18 Jan. 1870.

6. *Cong. Rec.* 45 (2), 25 Apr. 1878, 6 May 1878. *Cong. Rec.* 45 (3), 26 Feb. 1879. *Cong. Rec.* 46 (2), 5 Feb. 1880, 1 Apr. 1880, 6 Apr. 1880, 12 Apr. 1880, 13 Apr. 1880. *Cong. Rec.* 46 (3), 6 Feb. 1879, 18 Feb. 1879.

7. The Census Bureau did not officially become a bureau as such until the office was made permanent by an act of Congress in 1902. Up to this point the Bureau was referred to as the Census Office. [For clarity, this article refers to the Census Office as the Census Bureau, even though historically this was not the case.]

8. *Annual Report of Superintendent of Census*, *House Exec. Doc. No. 1*, 46(2), 15 Nov. 1879; "Introduction," *Compendium* (1883, xxxiv); *Cong. Rec.* 46(3), 18 Feb. 1879.

9. As Walker put it, "The work of census-taking is . . . exceptional and unique in its requirements . . . demanding a high degree of clerical capacity and a fitness to conceive and forcibly impart on the subordinate enumerator the many precise and delicate distinctions which are required properly to answer inquiries of the census schedules." *Annual Report of Superintendent of Census*, *House Exec. Doc. No. 1*, 45(3), 17 Jan. 1878.

10. Memorial to Congress, Ohio supervisors, National Archives, Record Group 46 (hereinafter NA RG), Senate 47A-H30.

11. *Congressional Globe* 46(3), 18 Feb. 1879, *St. Paul Pioneer Press* (hereinafter *SPPP*), "The Census of 1880," 6 June 1880, 11.3 (page eleven, column 3); *Washington Post* (hereinafter *WP*), "Carolina's Census," 10 Oct. 1880, 1.3.

12. 3 Mar. 1879, *Statutes at Large* 20.

13. Virginia supervisors complained, "their task has been at least doubly more onerous than could have reasonably been inferred from the census law," and their Indiana counterparts argued that "the work grew . . . to proportions far beyond what was contemplated when the service was accepted." Memorials to Congress, Virginia and Indiana supervisors, NA RG 233, 47th Congress, HR47–H24.1. For more examples, see Memorials to Congress from North Carolina, Ohio, and Illinois supervisors, NA RG 46, Senate 47A–H30; Memorials to Congress from Indiana and New Jersey supervisors, NA RG 233, 47th Congress, HR47–H24.1; *Additional Compensation for Supervisors of Census* 47 (1), House Report 1204, 3 May 1882.

14. Explained Walker, marshals "could not be made to feel that their tenure of office in any way depended upon his satisfaction with the manner in which their duties were discharged . . . The Department of the Interior would have to depend . . . simply upon the good-will of each of the sixty-four or sixty-five officers who are charged with this very onerous, thankless, and ill-requited duty. The department would be entirely powerless to bring any pressure to bear upon the marshal in the matter of his duty." *Provisions for Taking Tenth Census*, Interview with Superintendent of Census by Joint Committee 45 (3), *Senate Misc. Doc. No. 26*, 17 Dec. 1878.

15. Seaton (1883, xxxiv–xxxv); *Report of Superintendent of Census* 45(3), *House Exec. Doc. No. 1*, 17 Jan. 1878.

16. *Provisions for Taking Tenth Census*, "Interview with Superintendent of Census by Joint Committee," *Senate Misc. Doc. No. 26*, 46th Cong., 3d session, 17 Dec. 1878. See also *Report of Superintendent of Census* 45(3), *House Exec. Doc. No. 1*, 17 Jan. 1878; *Annual Report of Superintendent of Census* 46 (2), *House Exec. Doc. No. 1*, 15 Nov. 1879; *WP*, "Pricking a Political Bubble," 10 Sept. 1880, 1.2.

17. F. Walker, "Letter of Instructions to Supervisors of Census," 1 Feb. 1880, NA RG 128, Joint Committee of the Census, 46th Congress.

18. More broadly, this variation suggests that it is misleading to generalize about the quality and completeness of the enumeration on the basis of evidence from either the national office or a single community. While the national census office exercised far greater control and imposed more uniformity in practice in 1880 than in previous enumerations, the Tenth Census was still the end product of different practices carried out within 150 districts.

19. *Annual Report of Superintendent of Census* 46(2), *House Exec. Doc. No. 1*, 15 Nov. 1879. Such denials were presumably based on complaints from local residents. *Philadelphia Public Ledger* (hereinafter *PPL*), "Reports of Census Enumerators," 17 May 1880, 2.1; *Baltimore Sun* (hereinafter *BS*), "Census Enumerators," 31 May 1880, 1.8.

20. See, for example, oath of C. Johnson, 20 Feb. 1880, NA RG 48, Records of the Dept. of Interior, Appt. Division, Central Office, Appt. Papers, 1849–1907. *St. Louis Post Dispatch* (hereinafter *SLPD*), "Taking

the Census," 14 June 1880, 8.1; "An Act to Provide for Taking the Tenth and Subsequent Censuses," 3 Mar. 1879, *Statutes at Large* 20.

21. Walker reinforced this point in his directive on enumerator selection: "If it is badly done, in any district, the service will be discredited, the district will be disparaged in the result, and the Supervisor will not escape blame." In cities where local boosters' expectations of population increase were disappointed, even supervisors who had performed conscientiously found themselves reviled by newspapers and citizens' committees. Walker, "Letter of Instruction to Supervisors of Census," 1 Feb. 1880, NA RG 128, Joint Committee on the Census, 46th Cong. See, for example, the extensive and outraged reporting on the census in the *SLPD*, 18 June 1880 to 30 Oct. 1880.

22. Memorial to Congress, Illinois supervisors, NA RG 46, Senate 47A–H30.

23. During the 1870 census, Superintendent Walker had been frustrated by his lack of power to force marshals and assistant marshals to follow sound enumeration procedures. The appointment of supervisors directly answerable to the Census Bureau was an important structural change that opened the way for intensive oversight of fieldwork by the Washington office backed by formal sanctions.

24. "Introduction," *Compendium* (1883, xxxiv).

25. Mar. 1879, *Statutes at Large* 20. See also *PPL*, "The Census Enumerators Appointed by Mr. Steel," 31 May 1880, 1.7–8; *New Orleans Times Picayune* (hereinafter *NOTP*), "The Census Enumerators," 14 Aug. 1880, 2.2; *Minneapolis Tribune* (hereinafter *MT*), "The Men to Count," 24 May 1880, 7.1; *BS*, "Taking the Census," 28 May 1880, 2.1; *PPL*, "The Pay of Enumerator," 6 Aug. 1880, 3.8; *PPL*, "The Census Enumerators Appointed," 31 May 1880, 1.7–8; *PPL*, The Pay of Enumerators," 31 May 1880, 1.7–8; *New York Times* (hereinafter *NYT*), "The Census in New York," 22 May 1880, 2.3; Memorial to Congress, New Jersey supervisors, RG 233, H.R., 47th Cong., HR47–H24.1.

26. Mar. 1879, *Statutes at Large* 20.

27. Press reports indicate that the details of oversight varied across localities, ranging from direct supervisor review of schedules as in Baltimore, Atlanta, and New York City, to enumeration workshops in Philadelphia. *BS*, "Progress of the Census Taking," 3 June 1880, 1.8; *Atlanta Constitution* (hereinafter *AC*), "The Pending Census," 3 June 1880, 4.4; *NYT*, "Enumerators' Experiences in New York City," 5 June 1880, 10.1; *PPL*, "The Following Circular Has Been Mailed," 3 June 1880, 3.9; *PPL*, "Last Evening the Census Enumerators Met," 4 June 1880, 1.2; *PPL*, "Census Enumerators' Returns," 4 June 1880, 1.5; *PPL*, "How the Census Was Done," 1 July 1880, 2.2; *AC*, "The Census," 26 May 1880, 4.3; *SLPD*, "The Census," 6 Nov. 1880, 8.2–3; *SLPD*, "Aid the Census Takers," 8 Nov. 1880, 4.2; *Sacramento Daily Record* (hereinafter *SDR*), "The Census," 26 May 1880, 3.4; *PPL*, "Swearing in the Census Enumerators," 1 June 1880, 1.6.

28. Persons seeking to fill the post of enumerator were required to submit an application and "evidence of the capacity . . . to perform the duties contemplated" in the form of a test schedule, "write[ten] out in full." *Annual Report of the Director of the Twelfth Census* 56(2), *House Doc. No. 5*, 1 Nov. 1900, p. 293.

29. *Annual Report of the Director of the Twelfth Census* 56 (2), *House Doc. No. 5*, 1 Nov. 1900, p. 293.

30. U.S. Bureau of the Census (1910, 21).

31. *Report of the Director of the Twelfth Census* 56(2), *House Doc. No. 5*, 1 Nov. 1900, pp. 296–98.

32. *Enumerator's Street Book, Form 7–552*, NA RG 29, Entry 138, Box 1; File 703 (hereinafter NA RG 29, 138/1; 703).

33. *Report of the Director of the Twelfth Census* 56(2), *House Doc. 5*, 1 Nov. 1900, pp. 296–98.

34. U.S. Bureau of the Census (1911, 8).

35. U.S. Bureau of the Census (1910, 20–22).

36. U.S. Bureau of the Census (1911, 9)

37. Mar. 1879, Sect. 4, *Statutes at Large* 20; 1 Mar. 1889, Section 4, *Statutes at Large* 25; 3 Mar. 1899, Sect. 9, v. 30; 2 July 1909, *Statutes at Large* 35; Mar. 1919, Sect. 9, *Statutes at Large* 40; U.S. Bureau of the Census (1919a, 25).
The Bureau had often noted the need for speedy appointments, but in the case of the Fourteenth Census it was especially necessary; the enumeration date had been moved ahead six months from 1 June to 1 January.

38. The Bureau received approximately fifty-five hundred application requests, of which roughly twenty-two hundred were filled out and then returned. U.S. Bureau of the Census (1919a, 25).

39. Ibid.

40. NA RG 29, 212/179; 740.

41. U.S. Bureau of the Census (1929, 3); NA RG 29, 212/179; 740.

42. NA RG 29, 212/181; 25.

43. U.S. Bureau of the Census (1919b, 4).

44. U.S. Bureau of the Census (1919b, 3–4); NA RG 29, 212/182; 25; "Narrative to Be Used in Filling the Agriculture Schedule in the Test of Applicants for Appointment as Census Enumerator," NA RG 29, 138/1; 61.

45. NA RG 29, 212/182; 25.

46. U.S. Bureau of the Census (1940, 1–3).

47. Jenkins (1985).

48. Ibid.

49. U.S. Bureau of the Census (1940, 24–25); Jenkins (1985, 28). As incredible as it may seem, in 1930 the Census Bureau received test schedules from enumerators who filled out the exams using fictitious families or families of people they knew, rather than using the narrative printed at the bottom of the test schedule. For obvious reasons such applicants were deemed ineligible. The fact that these people were still allowed to be considered for the post if they agreed to retake the exam is testimony to the Census Bureau's degree of difficulty in obtaining any enumerators at all. NA RG 29, 212/179; 740. Other ineligibles screened out through this process included persons who were not U.S. citizens; current or retired federal employees; employees of the Census Bureau who were engaged in other Bureau inquiries; employees of states, municipalities, and other local government bodies prohibiting federal employment; or anyone under age 18.

50. Those who made the hiring preference fell into one or more of eight categories: war veterans and widows of war veterans (when equally qualified with others); crop reporters for the U.S. Department of Agriculture (if not USDA employees); retired farmers; graduates of or students at agricultural colleges; schoolteachers; town clerks, recorders, and other local officials; applicants who had at least a high school education; and, finally, applicants whose "appearance and manner indicated that they were suited for public contact." U.S. Bureau of the Census (1940, 25); Jenkins (1985).

51. See "Training Films for Enumerators" in *Preliminary Inventories*, National Archives (1964, 68); U.S. Bureau of the Census (1955, 16).

52. U.S. Bureau of the Census (1961).

53. Eckler (1972); Anderson (1988).

54. Riche (1993, 3); *Population Today* (1993, 10).

55. For discussion of alternative estimates of U.S. census undercounts and those who were likely to have been missed, see King and Magnuson (1995).

PART 2. Order Out of Chaos: The Integrated Public Use Microdata Series

General Design of the Integrated Public Use Microdata Series

Steven Ruggles, J. David Hacker, and Matthew Sobek

Public Use Microdata Samples (PUMS) of the U.S. census of population covering eleven census years between 1850 and 1990 are currently available or in preparation. Taken together, these microdata comprise our richest source of quantitative information on long-term changes in the American population. Because these samples were created at different times by different investigators, however, they have incompatible documentation and a wide variety of record layouts and coding schemes. These differences among the samples inhibit their use as a time series.

At the Social History Research Laboratory of the University of Minnesota, we are converting the series of public use samples into a single coherent form with uniform documentation. This article describes the general design of this Integrated Public Use Microdata Series (IPUMS).

Design of the Data Series

Detailed planning of the IPUMS was a major undertaking. Our goals were to maximize compatibility and ease of use, while retaining all significant information from the original samples. In some cases, these goals conflicted. We made most of the important design decisions early in the project in consultation with a National Advisory Panel, but we have continued to refine the design of the data series as we get feedback from early users of preliminary versions.[1] The following sections are intended to raise the most important design problems and to explain our general approach to them.

1. Record layout. Following conventional practice for public use samples, the IPUMS consists of numeric codes arranged in a column-format hierarchical structure. Variables common to the household as a whole—such as geographic indicators and housing questions—appear on a household record. Each household record is followed by a series of person records describing individual-level characteristics.

The design of the record layouts stresses column compatibility rather than compactness. In general, all variables available across multiple census years appear in the same columns in every year. When a variable is not available for a given year, the columns are filled with a missing data value. This means that the integrated versions of the public use samples will be substantially larger than the originals; we anticipate a record length in excess of 250 characters, almost twice the average of the existing samples. The great advantage of column compatibility is that it simplifies the construction of multiyear data files and minimizes the potential for error. In view of the rapid decline in the cost of mass data storage, it makes sense to focus on efficiency of use rather than efficiency of computing resources.

In the 1940 and 1950 census years, individuals falling on a designated "sample line" of the census form were asked an extra set of questions. The public use samples were constructed so that each household contains one sample-line individual, and the extra sample-line variables are provided on a separate sample-line record. The IPUMS eliminates the sample-line record by embedding the sample variables

in the person records. Individuals who were not asked the extra questions receive a missing data code for all sample-line variables.

2. Coding schemes. The existing PUMS employ different numeric classification systems in every census year, and reconciliation of these classifications was a major part of this project. For most variables, we found it impossible to construct a single uniform classification without an unacceptable loss of information. Some census years provide more detail than others; if we reduced all census years to their lowest common denominator, we would sharply reduce the power of the data series.

The household relationship variable, present in ten of the eleven existing PUMS, is a useful example for illustration. Most years have unique coding. For instance, an individual listed as a household head in the 1910 PUMS is represented by a code of "1000," but in the 1960 PUMS household heads are coded "0." The IPUMS employs consistent codes—household heads are coded "01" in all census years—eliminating this simple incompatibility. Of greater concern in creating a uniform coding scheme is the lack of detail available in the more modern censuses. The household relationship classification for the 1960–1970 census years consists of only fifteen categories. All other census years are more detailed; in fact, the 1910 census distinguishes 161 categories of household relationship. Tables 1 and 2 reproduce part of the original codebook pages describing the household relationship variable in the PUMS for 1910 and 1960. If we adopted the 1960 classification as a standard, we would lose the ability to distinguish such household relationships as nephews, aunts, farmhands, and domestic servants.

To avoid such problems and still maximize compatibility of coding systems, we designed composite coding systems for most variables: A general code, which provides the lowest common denominator available in all census years, and a detailed code, which provides additional information available in particular census years. The general codes usually constitute the first one or two columns of each variable and are for the most part entirely compatible. The detailed code, usually an additional one or two columns, provides the necessary detail to distinguish unique categories available in only one or more census years. This approach maximizes ease of use without losing information. It has been applied to all complex categorical classifications except for occupation, which is discussed at length later.

In the case of household relationship, for example, the first two digits are essentially equivalent to the fifteen-category coding system of the 1960 PUMS. The IPUMS codebook pages for family relationship are shown in table 3. As the table illustrates, relationship is listed under two variables in the codebook: general relationship and detailed relationship. The record layout noted in the header, however, indicates that the column locations—person record 21-22 for general relationship and person record 21-24 for detailed relationship—overlap; the detailed relationship shares the first two characters of the general code. [A space in the codebook separating the general and detailed code is shown for detailed relationship to help improve readability. No blank spaces exist in the dataset.] The numbers under the census years indicate the case count in each year: where no numbers appear, the category is not available in that year. Researchers interested in a strictly compatible relationship variable across all PUMS would use only the first two columns, separately listed in the codebook as general relationship. Researchers interested in a category not available in all censuses would use all four columns, listed in the codebook as detailed relationship. The classification for stepchild, for instance, is not available in the 1960, 1970, or 1980 PUMS, necessitating that it be assigned a detail code under the general code for child, the classification that was assigned to stepchildren during these census years.

TABLE 1
1910 Relationship Codes (Partial Listing)

P06 REL = Relationship to head
Columns 14–17
Width 4

Value	Description	No. of individuals	%
–3	Unknown	12	0.00
–2	Illegible	44	0.01
–1	Blank	910	0.25
1000	Head	80,589	22.00
1201	Husband	26	0.01
1202	Wife	63,773	17.41
1300	Child	3	0.00
1301	Son	82,168	22.44
1302	Daughter	78,407	21.41
1310	Stepchild	9	0.00
1311	Stepson	1,429	0.39
1313	Stepdaughter	1,279	0.35
1320	Adopted child	21	0.01
1321	Adopted son	187	0.05
1322	Adopted daughter	219	0.06
1331	Son-in-law	1,020	0.28
1332	Daughter-in-law	840	0.23
1341	Stepson-in-law	8	0.00
1342	Stepdaughter-in-law	10	0.00
:			
2301	Nephew	1,430	0.39
2302	Niece	1,485	0.41
:			
9890	Unclassifiable	113	0.03
Total		366,239	100.00

Source: Strong et al. 1989.
Note: To conserve space, only a portion of the 161 relationship codes in the 1910 PUMS are listed. Common relationships not shown above include aunt, uncle, and servant.

TABLE 2
1960 Relationship Codes

Character	Item and data descriptor name	Code	Description of codes
P1	Basic relationship	0	Head of household
		1	Wife of head
	HEADRELA	2	Son or daugther of head
		3	Other relative of head
	Dict.: 79,80	4	Roomer, boarder, or lodger
		5	Patient or inmate
		6	Other not related to head
P2	Detailed relationship of persons in households	—	Head, wife, or child not in subfamily, or G.Q.
		0	Son or daughter or son-in-law or daughter-in-law in subfamily
	HEADRELB	1	Grandson or granddaughter
		2	Father or mother or stepparent
	Dict.: 80	3	Father-in-law or mother-in-law
		4	Brother or sister or stepbrother or stepsister
		5	Brother or sister-in-law
		6	Other relative—aunt, cousin, etc.
		7	Partner or friend
		8	Roomer, boarder, or lodger
		9	Resident employee

Source: U.S. Bureau of the Census 1973a.

The conversion of relationship coding in individual PUMS to the IPUMS format was accomplished with the aid of a translation table, a look-up table for our dataset construction program. In addition to facilitating programming, translation tables provide documentation of the various PUMS, as well as our construction decisions, and are included in the IPUMS documentation. A partial listing of the translation table for relationship is shown in table 4. In addition to noting the column locations in which relationship can be found in the original PUMS, the translation table indicates the coding used in each PUMS: where no numbers appear, the category is not available in that year.

3. Design of occupational and industrial classifications. Occupation is the most complex variable collected by the census. It is also among the most important census variables for analysis of long-term social change because the early census years provide few alternative indicators of socioeconomic status or labor-force participation. The Census Bureau has modified its classification systems every decade, so all comparisons of occupation and industry require extensive reconciliation of codes. There are nine different occupational classification systems consisting of between 285 and 550 categories each. Although a complete reconciliation of these coding schemes is impossible, we are providing variables that maximize the potential for consistent comparisons of occupational status. These variables are described elsewhere in this issue in Matthew Sobek's article (see pp. 47–51).

4. Design of group quarters classification. Variations in the treatment of group quarters among the public use samples can introduce incompatibilities in a variety of variables. All the samples incorporate provisions for individual-level sampling of large units such as institutions and boarding houses, but the criteria for individual-level sampling have varied from year to year. Following current Census Bureau practice, we refer to persons sampled at the individual—rather than the household—level as residents of group quarters. For the census years 1940 through 1970, group quarters are defined as units containing five or more persons unrelated to the household head. In the samples for 1850, 1880, and 1920, by contrast, units must be larger than thirty before they are sampled as group quarters. The other census years have definitions that fall between these extremes.

The variation in the definition of group quarters has important implications for compatibility of the samples. Since the residents of group quarters are sampled at the individual level, they cannot be evaluated in the context of other residents of their living units. Moreover, many census questions for the period since 1940 were not asked of residents of group quarters.

The 1970 sampling rules constitute a lowest common denominator for definitions of group quarters. Fortunately,

TABLE 3
IPUMS Relationship Codes (Partial Listing)

	Code	Census year								
		1880	1900	1910	1940	1950	1960	1970	1980	1990
P21-22 RELATE—G = Relationship—general										
Head/householder	01	101724	21333	80589	350354	443719	530199	634408	767350	918782
Spouse	02	81446	16682	63779	267997	375009	396000	437135	466948	532985
Child	03	246379	47027	163722	540738	854306	697452	783186	726472	785662
Child-in-law	04	2394	466	1878	12275	22678	5835	5002	4254	3801
Parent	05	385	944	3053	12315	18448	12843	12440	13939	16593
Parent-in-law	06	2405	568	2433	10654	19873	12905	10241	6005	3686
Sibling	07	6055	1336	4877	17231	19839	14398	15043	19541	23376
Sibling-in-law	08	2821	688	2975	10318	15798	7264	5737	4365	316
Grandchild	09	28	1578	5375	26400	56257	26366	25623	25477	42869
Other relative	10	5556	112	4248	15257	25977	21150	13248	14991	19800
Roomers, boarders, or lodgers	11	16486	4967	21131	34041	42163	21150	20782	14662	13862
Other nonrelative in household	12	14588	2275	8082	13472	10719	5313	12065	43233	79188
Group quarters unspecified	13	218	643	563	27849	12592	30285	33389	31383	27120
Institutional inmates	14	1476	686	2435	12831	4820	18943	21334	23769	25072
P21-24 RELATE—D = Relationship—detailed										
Head/householder	01 01	101724	21333	80589	350354	443719	530199	634408	767350	918782
Spouse	02 01	81428	16679	63799	267997	375009	396000	437135	466948	532985
2d/3d wife (polygamous)	02 02	18	3							
Child										
Including adopted, stepchild	03 01	241802	46169	160578	531847	837706	697452	783186	726472	745382
Adopted child	03 02	75	103	427						
Stepchild	03 03	3820	755	2717	8891	16600				40280
Child-in-law	04 01	2353	466	1860	12275	22678	5835	5002	4254	3801
Stepchild-in-law	04 02	41		18						
:	:									
Other relatives										
Grandparent	10 11	196	27	112	519	940			705	404
Stepgrandparent	10 12	1	1	2						
Grandparent-in-law	10 13	13	2	15						
Aunt or uncle	10 21	423	99	385	1630	2792			1277	891
Aunt, uncle-in-law	10 22	7	2	16						
Nephew, niece	10 31	4058	822	2915	10317	15874			7999	6744
Nephew, niece-in-law	10 32	24	4	29						
Step/adopted nephew/niece	10 33		1	5						
:	:									
Institutional inmates	14 01	350	478	1581	12831	4820	18943	21334	23769	25072
Convicts	14 11	450	99	334						
Paupers	14 12	146	9	3						
Inmate, insane asylum	14 13	15	9							
Hospital patient	14 14	238	17	347						
Orphan	14 15	198	69	60						
Religious institution	14 16	79	5	110						

Source: Ruggles et al. 1993.

sufficient information is available in all census years to determine what the sampling unit of each case would have been under the 1970 rules. Thus, it is possible in all census years to suppress information on household composition, family relationships, and other variables not asked of residents of group quarters if that information would not have been available under the 1970 rules. Unfortunately, the 1970 rules are somewhat inappropriate for the early twentieth century, when many households included five or more boarders or servants and a high percentage of the population resided in secondary families. Our experience has shown that many users require the greater precision available in the early census years. We have therefore retained all available household and family information but have constructed an individual-level variable indicating the sampling unit under 1970 rules. The documentation provides full instructions on the use of this variable to convert each family and household variable into fully consistent form.

The differences among the samples in treatment of group quarters, together with other variations in sample design, have important consequences for precision of the samples. These differences among the samples are described in the

TABLE 4
IPUMS Relationship Translation Table (Partial Listing)

P21-24 RELATE = Relationship

PUMS	Columns	PUMS	Columns
1900	P 09 11	1960	P 01 02
1910	P 14 16	1970	P 01 02
1940	P 11 12	1980	P 02 04
1950	P 16 17	1990	P 09 10

	IPUMS code	Original PUMS coding								
		1880	1900	1910	1940	1950	1960	1970	1980	1990
Head/householder	01 01	100	100	100	01	01	0–	0–	000	00
	01 01						00	00		
Spouse	02 01	120	120	120	02	02	1–	1–	010	01
Husband, not head	02 01	140	140							
2d/3d wife (polygamous)	02 02	121	129							
Child										
Including adopted, stepchild	03 01	130	130	130	03	03	2–	2–	020	02
	03 01						20	20		
Adopted child	03 02	132	132	132						
Stepchild	03 03	131	131	131	04	04				03
Child-in-law	04 01	133	133	133	05	05	30	30	051	07
Stepchild-in-law	04 02	134		134						
:	:									
Other relatives										
Grandparent	10 11	260	260	260	09	09			056	*
Stepgrandparent	10 12	261	261	261						
Grandparent-in-law	10 13	263	263	263						
Aunt or uncle	10 21	250	250	250	12	12			057	*
	10 21			254						
Aunt, uncle-in-law	10 22	253	253	253						
Nephew, niece	10 31	230	230	230	13	13			055	*
Nephew, niece-in-law	10 32	233	233	233						
Step/adopted nephew/niece	10 33			231						
	10 33		232							
:	:									
Institutional inmates	14 01	700	700	800	99	99	5–	5–	100	12
Convicts	14 11	720	720	810						
Paupers	14 12	730	730	912						
	14 12		510							
	14 12	731								
Inmate, insane asylum	14 13	740	740							
Hospital patient	14 14	750	750	820						
Orphan	14 15	760	760	830						
Religious institution	14 16	790	790	860						

Source: Ruggles et al. 1993.
Note: An asterisk in the original PUMS coding indicates the need for special programming. In the case of the 1990 detailed relationship variable, the general and detailed codes are located in noncontiguous columns, necessitating a small modification in our translation program.

Ruggles article, "Sample Designs and Sampling Errors" on pages 40–46 in this issue.

5. Design of constructed variables on family interrelationships. Individual-level constructed variables describing interrelationships among family members are not consistently available in the original PUMS. Such variables make it possible for users to create specialized measures of living arrangements tailored to their specific research topics, such as living arrangements of the elderly or of single parents. These indicators also facilitate the construction of specialized own-child fertility measures and measures of marriage characteristics. The IPUMS contains three pointer variables that give the location within the household of each individual's spouse, mother, and father. The pointer variables allow users to easily attach characteristics of these kin, and users find them to be convenient tools for the construction of measures of fertility and coresidence. The data series also includes several of the most commonly requested variables on own children: number of own children, number of own children under 5 years old, age of eldest own child, and age of youngest own child. The Ruggles article "Family Inter-

relationships" later in this issue (see pp. 52–58) describes these variables in detail.

6. Design of geographic codes. 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 suppressed in the public use samples of the period 1940–1990. In 1960 and 1970, places with fewer than 250,000 inhabitants were not identified; for 1940–1950 and 1970–1990, the threshold for identification is 100,000. Within these constraints, the classification systems for identifying places within states varies considerably. The 1940 and 1950 samples provide State Economic Area (SEA), which is a system for coding county groups, and Standard Metropolitan Area (SMA). No information on urban/rural residence is given for those years. In the 1960 Public Use Sample, no geographic locations below the state level are available, but there is a variable on urban/rural residence. In both 1970 and 1980, the Census Bureau released three versions of the public use samples containing alternate geographic variables, and there are two versions for the 1990 census. In spite of this, there remain significant problems of compatibility in the geographic codes for these three years. Both 1970 and 1980, however, do identify Standard Metropolitan Statistical Area (SMSA), and the 1990 census identifies the closely comparable system of Metropolitan Area (MA). The definition of SMSA differs slightly from the earlier SMA, but they can be viewed as compatible for many applications. In addition, some county groups can be consistently identified from 1970 to 1990, and all three years identify urban/rural residence. The 1940–1950 and 1980–1990 census years also include sufficient information to identify consistently the residents of sixty-one of the largest cities.

We cannot do much about the geographic incompatibilities of the 1940–1980 census years, other than imposing consistent numeric codes for SMA and SMSA. What we can do is construct variables for the early census years that are compatible with the later public use samples. The samples for 1850–1920 provide full information on county, city, and enumeration district. It is therefore possible to construct the SEA variables used in recent census years, although the correspondence is not precise because of boundary changes and creation of new counties. We are also creating the closest possible analogs of SMA and urban/rural residence for the early census years. Although the definitions of these variables depend in part on commuting ties, telephone calls, and other measures of integration and metropolitan character that are not available for the earlier census years, reasonable approximations can be constructed on the basis of other available information.

7. Design of all other coding systems and constructed variables. The classification issues previously discussed are the most problematic ones, but there are a wide variety of other variables that required significant work. For example, the development of a consistent scheme for classification of countries of birth and parental birth has been complicated by dramatic boundary changes in many parts of the world since the mid-nineteenth century. Other complex classifications include mother tongue, institution type, and ancestry. Even such apparently straightforward classifications as income, employment status, and education required manipulation to make them compatible across census years.

To aid file handling, the data series also includes a basic set of constructed variables, such as census year, sample number, record type, number of person records in sample unit, household sequence number, and person sequence number within households.

Documentation

The creation of integrated documentation is a top priority of this project. By comparison with the usual standards of social science research, the existing documentation of the public use samples is quite good, but it still has significant limitations. Among the eleven existing samples, only two were ever documented as fully as had been originally planned. Indeed, in two cases—the 1940 and 1950 PUMS, which fail to document necessary weighting procedures—it is impossible to use the samples correctly if one relies entirely on the documentation supplied with the data.

The documentation is particularly awkward if one uses the public use samples as a time series. With few exceptions, the documentation for each census year is organized differently. The combined documentation adds up to some three thousand pages, and there are usually no indexes. Simply learning how to look up things in each census year requires a substantial investment of time. Moreover, with the exception of the most recent census years, the discussions of comparability issues range from inadequate to nonexistent.

We plan a three-volume set of documentation, consisting of a general user's guide, a volume on comparability issues and procedural histories, and a volume on technical characteristics and error estimation. Each volume will be about five hundred pages long.

The user's guide will contain the essential information for routine use of the data series. It will include a general description of the public use samples, guidelines for use of the data series, record contents descriptions, a glossary of terms, and a brief summary of sample designs and error estimates.

The volume on procedural histories and comparability issues will provide a comprehensive treatment of changes in the census that affect comparability of the public use samples. We will include capsule procedural histories for all census years and complete enumerator instructions organized by variable. We will focus especially on problems of comparability that stem from differences in enumeration

procedures and on changes in post-enumeration editing and processing. For the period since 1950, our principal source is to be the official procedural histories prepared by the Census Bureau (U.S. Bureau of the Census 1955, 1966, 1976, 1986–1989). In the case of the 1940 census, we are using a new procedural history created in conjunction with the 1940 PUMS (Jenkins 1987). For the early census years, where no official procedural histories exist, we are relying on the primary research of project assistant Diana Magnuson (forthcoming).

The third volume of documentation will contain additional detail on many topics covered briefly in the user's guide, including data on verification results, approximate standard errors, and allocation statistics. We will also provide full documentation of the conversion of the original public use samples into their integrated format, with particular attention to sources of imprecision in the occupational and geographic codes. Finally, we will include a set of frequency distributions for key variables.

Release of Data

A preliminary beta-test version of the data series was released in November 1993 and updated in November 1994. The final version will be released through the Inter-university Consortium for Political and Social Research (ICPSR) and the National Archives in November 1995. We intend to release the data in several different formats. The public use samples for the period since 1960 are divided into files according to geographic area, whereas those for the earlier census years are broken into nationally representative subsamples. Geographic organization is most convenient for state and local policy analysts and others with an interest in particular regions, but for most academic investigators the nationally representative subsamples are more useful. We will therefore make the IPUMS available in both forms.

We also plan a compact edition of the data series. This version will maximize comparability at the expense of information. It will include only the common-format component of all composite variables and will eliminate all variables not available in multiple census years. In addition, we will suppress information on household composition and family relationships if it would not have been available in all census years. The compact edition will be considerably simpler and smaller than the main version of the data series and will thus be more efficient for users who do not require fine detail. Finally, we plan to release merged data files containing data from all nine census years. These files will be designed primarily for teaching purposes and for exploratory data analysis. They will contain a small representative sample of records drawn from the compact edition of each census year.

Conclusion

Producing standardized variables for eleven PUMS containing numerous questions and deviations in categorization, wording, and universe involves countless decisions similar to those previously described for family relationship. In many cases we were not able to produce strictly compatible variables. In such instances the user is instructed to pay careful attention to the IPUMS documentation. For the great majority of variables, however, we believe that we have succeeded in creating an integrated data series that eliminates the confusion surrounding the comparability of PUMS.

The IPUMS promises to greatly enhance the manageability and power of the historical census datasets. It will make the national census files more readily accessible to a broad range of users and allow temporal analyses of topics such as fertility behavior, urbanization, immigration, household composition, and occupational structure. The extensive combined documentation will make researchers more cognizant of subtle and not-so-subtle variable changes that may influence their analytical design and interpretation of results. Perhaps most important, the availability of consistent coding schemes and uniform documentation will reduce the potential for error and make results more easily reproducible. Our hope is that the IPUMS will be used by many investigators in the years ahead.

NOTES

1. The members of the advisory panel included Stewart Tolnay, Miriam L. King, Michael R. Haines, Margo Anderson, and Myron P. Gutmann.

PART 2. Order Out of Chaos: The Integrated Public Use Microdata Series

Sample Designs and Sampling Errors

Steven Ruggles

The Public Use Microdata Samples (PUMS) of the U.S. census are based on a variety of different sample designs. These variations in sample design have significant implications for the precision of sample estimates. Estimates derived from any sample are subject to sampling variability, which is usually measured as the standard error. The standard error of a sample statistic estimates the variation of that statistic across many similar samples drawn from the same population. Approximately two-thirds of random samples will produce estimates within one standard error of the full population, and approximately 95 percent of samples produce estimates within two standard errors. Standard errors depend on both sample size and sample design. This article outlines the aspects of sample design that influence sample precision in the PUMS, provides estimates of the resulting differences in standard errors, and discusses strategies for obtaining realistic estimates of statistical significance. The Appendix briefly describes the sample designs used to create each sample.

Clustering

All the PUMS are cluster samples. Most information of interest in the census concerns individual characteristics, such as age, race, sex, income, education, and so on. But the PUMS are not individual-level samples; instead, they are samples of households or dwellings. The information about individuals is gathered household by household because many important topics of analysis—such as fertility, household composition, and nuptiality—require information about multiple individuals within the same household. The number of independent observations in each census file is the number of households or dwellings, not the number of individuals.

Some individual characteristics—such as ancestry—are highly correlated within households. If one household member is Chinese, for example, the odds are high that other household members will also be Chinese. Suppose we wish to estimate the standard error for the proportion of the population of Chinese ancestry. If we had the sort of sample generally assumed by statistics textbooks—an independent random sample of all individuals in the population—the standard error for Chinese ancestry would be inversely proportional to the square root of the number of individuals in the sample. But because the PUMS are cluster samples, and Chinese ancestry is highly correlated within clusters, the usual method for calculating standard errors would overestimate sample precision. A better estimate of standard error would be obtained by substituting the number of households for the number of individuals when making the calculation.

Standard errors in cluster samples depend on both the number of clusters sampled and on the homogeneity of variables within clusters. Calculation of standard errors for cluster samples is complicated (Hansen, Hurwitz, and Madow 1953; Kish 1965). 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 clusters rather than the number of individuals. For variables that are heterogeneous within clusters, such as age and sex, clustering may have little or no effect on sample precision.

The impact of clustering therefore varies from variable to variable. It also varies from census year to census year. The homogeneity of particular characteristics within clusters can change over time. For example, with increasing ethnic intermarriage, the homogeneity of ethnicity within households would be expected to decline.

The size of clusters also differs across census years. The larger the average size of clusters, the smaller will be the number of independent observations. With the fall in fertility, the decline of boarding and extended family structure, and the rise of the primary individual, household size has fallen dramatically over the past century; thus the PUMS files for recent years have far smaller clusters, on average, than those for earlier years.

Changes in cluster size occur not only because of change over time in household size but also because of differences among the samples in treatment of group quarters. Group quarters are large units such as institutions, boarding houses, and college dormitories. To maximize sample precision, such units are sampled at the individual level in all PUMS. Thus, for example, instead of treating a one-thousand-inmate prison as a single sample unit, it is sampled as if it were one thousand one-person households. This procedure multiplies manyfold the number of independent observations for persons in large units. Unfortunately, the criteria for designating group quarters vary considerably among the samples.

Table 1 summarizes the definition of group quarters in each census year. In general, the census years 1940 through 1970 have the broadest definition of group quarters: all persons in units with five or more persons unrelated to the household head are sampled as individuals. Thus, for example, a wealthy family with five coresident servants would be treated as group quarters, and each member of the unit—including the primary family—would be sampled as if he or she resided in a separate one-person household. This inclusive definition of group quarters has the effect of minimizing the size of clusters. In 1850, 1880, and 1920, on the other hand, a unit may have up to thirty members, related or unrelated, before it is sampled as group quarters. In a unit with over thirty members, each related group is sampled jointly in order to preserve all coresident family relationships. This is a minimal definition of group quarters that maximizes the size of clusters; sample precision was reduced in order to preserve information about family and household interrelationships. The other census years fall between these extremes; further details appear in the Appendix.

There is one additional feature of the sample designs that affects the size of clusters. Most public use files are samples of households, individuals within households, and group quarters.[1] For the samples of 1850, 1880, and 1920, however, we added another level of hierarchy. When we encountered a multihousehold dwelling with thirty or fewer residents, we sampled at the level of the dwelling instead of the level of the household by including in the sample all households within the dwelling. This provides important additional information, since many dwellings contained two interrelated households. At the same time, sampling by dwellings resulted in some sacrifice of sample precision, since it increased the average size of clusters.

Stratification

The sample designs for the public use census files all incorporate implicit or explicit stratification. Stratification has the opposite effect of clustering: it increases the precision of sample estimates. In some cases, the positive effects of stratification outweigh the adverse effects of clustering, so the PUMS sample designs can actually yield smaller standard errors than would be obtained through a simple random sample of similar size.

Stratification involves dividing the population into strata based on key characteristics and then sampling separately from each stratum. This strategy ensures that each stratum is proportionately represented in the final sample. The method increases precision not only for those characteristics that are explicitly stratified but also for any other characteristics correlated with them.

TABLE 1
Group Quarters Definitions

Census years	Definition
1850, 1880, 1920	Units of size 31 or more; related groups within group quarters sampled jointly
1900	See Appendix
1910	Units with 21 or more members unrelated to household head; related groups within group quarters sampled individually
1940, 1950, 1960, 1970	Institutions and other units with 5 or more members unrelated to household head; related groups within group quarters sampled individually
1980, 1990	Institutions and other units with 10 or more members unrelated to householder; related groups within group quarters sampled individually

Note: The 1900 sample is excluded because it lacks a group quarters concept; instead, all persons unrelated to the household head except employees are sampled as individuals or related groups. This strategy makes the 1900 sample incompatible with the other PUMS for some applications; see Appendix.

The PUMS for the four most recent census years are all based on elaborate stratification schemes, which are described in some detail in the Appendix. To take an example, the 1960 census divided the population into 38 strata, based on household size, homeownership, race, and group quarters residence, and systematically selected households from each stratum for inclusion in the public use file. Sample precision was further enhanced by a selection scheme that ensured even coverage within every geographic area. The stratification grew more elaborate in every subsequent census year; by 1990, the PUMS was selected from 1,049 strata.

Such elaborate sample designs have not been feasible for the historical PUMS files covering the period 1850 to 1950. Unlike the more recent census years, these census files were not drawn from an existing machine-readable source; instead, they were entered by hand from microfilm of the original enumerators' manuscripts. Most individual and household characteristics were unknown before the cases were entered into the sample and could therefore not be used as an efficient basis for selecting the samples.

One key characteristic was available for every case prior to data entry, however. The microfilm reels containing the census manuscripts are organized geographically and, within microfilm reels, the sequence of census pages is also organized geographically. Each of the historical PUMS first divided the raw data into individual census pages, groups of pages, or individual microfilm reels, and then sampled independently from each of these strata. The result is a significantly more even geographical distribution of cases than would be expected from a true random sample. This dramatically improves the precision of the geographic variables, such as region and urban residence. It also indirectly improves the precision of variables highly correlated with geography (e.g., race, ethnicity, education, occupation, farm residence, and homeownership).

Estimating Sampling Errors in the PUMS

For a true random sample in which each case is an independent observation, it is a simple matter to estimate the probability of error due to sampling even before the sample is drawn. For samples that are simultaneously clustered and stratified, it is much more difficult. Analytic estimates based on theory are possible, but they are usually oversimplifications of the problem even though the calculations involved are inevitably complex.

Once the samples are complete, however, it is fairly easy to develop empirical estimates of standard errors. Standard errors are simply estimates of the standard deviation of a statistic over all possible samples of a population. The PUMS are large enough that we can divide them into many subsample replicates and directly measure the distribution of a statistic across the subsamples.

Table 2 shows estimated "design factors" for selected variables in each PUMS file from 1880 to 1980. These factors were estimated by dividing each PUMS into fifty randomly selected subsample replicates, calculating the standard deviation of the expected value of each variable across the fifty subsamples, and dividing the result by the standard error predicted by statistical theory for a simple random sample of the same size as each subsample.[2]

The design factors represent the ratio of observed standard errors for a variable to the standard errors that would be obtained from a simple random sample of the same size. Thus, a design factor of 1.0 means that the effects of stratification and clustering on sample precision cancel one another out. If the design factor is 1.0, a standard statistical package like SPSS or SAS—or a standard statistics textbook—would produce reliable significance statistics. A design factor of 2.0 means that the empirically observed standard errors are twice as great as would be predicted for a simple random sample. For such variables, a statistical package would overestimate statistical significance. Conversely, a design factor of 0.5 means that the PUMS file is twice as precise as would by predicted by standard statistical tests.

Only a few variables have design factors that usually exceed 1.0 by a wide margin. The most dramatic case is race, where the design factor exceeds 2.0 in each of the census years before 1960. This reflects the impact of clustering, since households have historically been extremely homogeneous with respect to race. The stratification schemes adopted since 1960 have reduced the design factor for race. Birthplace, language, and citizenship also tend to be homogeneous within clusters and have relatively high design factors. A falling design factor for school attendance is probably attributable to declining fertility: as fewer and fewer households have many school-age children simultaneously, the potential for clustering has diminished.

The design factors presented in table 2 are only valid for analyses of all individuals in the nation as a whole; the results could be significantly different for any population subgroup. Moreover, particular categories of variables may have design factors markedly different from the variable as a whole. For example, the categories of "head" and "wife" on the relationship variable have uniformly low design factors; because each household ordinarily contains only one head and no more than one wife, there is no potential for homogeneity within households. By contrast, however, the design factor for the relationship category "child" is quite high because children tend to occur in combination.

Most common individual-level analyses have lower design factors than are indicated in table 2 because analysts tend to choose population subgroups that minimize the impact of clustering. For example, fertility studies most frequently focus on married women aged 15 to 49. Since the great majority of sample clusters contain only one such individual, the impact of clustering is trivial. Thus, fertility studies almost inevitably have design factors at or below 1.0, and any significance statistics will therefore be conser-

TABLE 2
Design Factors for Selected Variables, 1880–1980

Variable	1880	1900	1910	1940	1950	1960	1970	1980
Relationship	1.0	0.9	1.0	0.9	0.9	0.5	0.4	0.6
Age	1.1	1.1	1.1	1.2	1.1	1.0	0.9	1.1
Sex	0.8	0.9	0.8	0.8	0.7	0.7	0.8	0.7
Marital status	1.0	1.0	0.9	1.0	1.0	0.7	0.5	0.8
Race	2.2	2.1	2.1	2.1	2.1	0.5	1.4	1.3
Occupation	1.2	1.1	1.2	1.1	0.8	0.9	1.0	1.0
Place of birth	1.3	1.3	1.5	1.2	1.7	1.1	1.1	1.1
Language	—	—	1.7	0.8[a]	—	1.3	1.6	1.4
Citizenship	—	1.1	1.3	1.2	1.3	0.9	1.2	1.3
School attendance	1.3	1.3	1.2	1.2	0.8[a]	0.9	0.9	1.0
Home ownership	—	0.9	0.9	0.9	—	0.1	0.4	0.9
Urban/rural residence	0.8	0.7	0.7	—	—	1.0	0.8	—
Farm residence	0.9	1.0	0.9	0.9	1.1	1.0	1.0	1.1

[a]Sample-line variable.

TABLE 3
Design Factors for School Attendance, 1880 PUMS

Population subgroup	Factor
Total population	1.33
Girls aged 10–14	1.08
Randomly selected individual aged 5–19	0.99

vative. Likewise, investigations of such topics as the living arrangements of elderly women, the occupational status of young men, or the education of never-married adult women should generally yield precision at least as high as would be obtained from a simple random sample of the same size.

Researchers can usually set up their analyses to avoid high design factors. For example, consider the design factors shown in table 3 for school attendance in 1880. For the population as a whole, the design factor is 1.33, indicating that the standard error for school attendance in the PUMS would be 33 percent larger than in a simple random sample. If we restrict the analysis by age and sex and look at the school attendance of girls aged 10–14, the design factor improves markedly to 1.08—because most clusters include only one girl aged 10 to 14. Finally, if we randomly select one school-age child from each household, the design factor falls below 1.0. If we design our studies so we rarely or never use more than one individual per household, it is possible to avoid the effects of clustering altogether.

Users can therefore take simple steps to reduce the impact of clustering. Especially when doing analyses of children, boarders, and other groups likely to appear multiple times in the same household, researchers should develop strategies to eliminate the redundant cases. Instead of assessing the characteristics of all children, for example, one can look at eldest children, or youngest children, or children of a particular age, or a randomly selected child from each household.

In practice, users of the PUMS never take the trouble to estimate true standard errors, mainly because the methods for doing so are quite cumbersome. Instead, users universally accept the significance statistics generated by their statistical packages. On the whole, the results presented here are reassuring: most of the analyses done by users of the PUMS probably have design factors close to 1.0 or lower, which means that the estimates of the packages are not too far off. If users are aware of the dangers of clustering and design their studies to minimize it, they can safely use statistical procedures designed for simple random samples.

APPENDIX

Summaries of IPUMS Sample Designs

For all census years from 1850 to 1950 (except for 1890, which does not survive) the population schedules are preserved on microfilm at the National Archives in Washington, D.C. In each year, the microfilm reels and the schedules within reels are organized geographically: alphabetically by state, within states alphabetically by county, and within counties numerically by enumeration district. For census years since 1960, the census schedules exist in fully machine-readable form.

The sample designs for all years are constrained by the available units of enumeration. In the census years prior to 1940, all individuals were assigned to a "family." The definition of the census family varied only slightly from census to census between 1850 and 1920. A *census family* was an individual or group of individuals living together in the same dwelling place. Census instructions defined a *dwelling place* as any occupied structure. Two or more families could reside in a single dwelling place, provided they occupied separate parts of the house and their housekeeping was separate. However, all the permanent occupants of hotels, institutions, and military barracks constituted single families. Census enumerators likewise counted boarders, lodgers, and servants as part of the family occupying the dwelling place where they slept, regardless of their housekeeping arrangements. In 1940, the basic unit of enumeration shift-

ed from the census family to households and quasi-households. A *household* consisted of the group of persons occupying a dwelling place or part of a dwelling place with either separate cooking equipment or an outside entrance. The maximum number of boarders and lodgers in a household was ten; 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. The 1950 census was similar, except that quasi-households included all units with five or more persons unrelated to the household head. In the years since 1950, the term *group quarters* has been substituted for the term quasi-household, and there have been minor variations in the criteria used to distinguish separate households within the same structure. In 1980, the number of unrelated persons required for group quarters classification was raised from five to ten. Further details on the comparability of census enumeration units can be found in Ruggles (1991a) and Smith (1992).

For the census years before 1950, the sample units contained in the PUMS are subsets of the original enumeration units. The following sections describe the sample units used for each census year. In addition, each section briefly describes the procedures used to select cases for inclusion in the PUMS.

1850. The manuscript census of the 1850 free population consists of roughly 560,000 census pages recorded on 976 reels of microfilm. Each census page has eighty-four lines, and the information pertaining to each individual appears on a separate line.

On each microfilm reel, we selected census pages systematically, ordinarily at intervals of six pages. On each selected census page, we randomly selected one line, designated the sample point. Any valid sample unit beginning at the sample point or within four subsequent lines was included in the sample. This procedure yields a 1-in-100 sample with equal probabilities of inclusion for all individuals and households. Valid sample units are defined as follows:

- **Dwellings**: dwelling units under size thirty-one, with or without multiple households.
- **Households:** census families under size thirty-one in dwellings size thirty-one or over.
- **Related groups in group quarters**: groups related by blood or marriage in census families over size thirty. Family relationships inferred from surnames.
- **Individuals in group quarters**: unrelated individuals in census families over size thirty.

1880. The manuscript census for 1880 consists of about 600,000 enumeration pages with one hundred lines per page. These records are contained on 1,454 reels of microfilm.

We randomly selected one line on each page and designated this as the sample point. Sample units were included only if they began at the designated sample point. Valid sample units were defined the same as in 1850, except that related groups in group quarters could be identified through the family relationship variable as well as by surname. This procedure yields a 1-in-100 sample with equal probabilities of inclusion for all individuals and households.

1900. The 1900 manuscript census consists of some 900,000 census pages contained on 1,850 microfilm reels. Each page contains one hundred lines.

The measured length of each reel was used to estimate the number of census lines on the reel. One in 750 of these lines were randomly selected and designated as sample points. Cases were entered if a sample point fell on the first individual in a valid sample unit. Sample units were defined as follows:

- **Family:** head of census family, all persons related to the head, and all coresident employees of the head (servants and domestic farm workers).
- **Institutional employees:** all individuals residing at and employed at institutions were sampled as a single unit.
- **Related nonfamily groups:** all persons who were neither related to nor employed by a family head, and who were not employees of institutions but who were related to others in a census family, were sampled as a related group. These sample units were primarily composed of boarders and lodgers but, in theory, could also include such persons as inmates in institutions and military personnel.
- **Nonfamily individuals:** boarders, lodgers, inmates, military personnel, and all others residing without any family were sampled as individuals.

The documentation for the 1900 sample refers to the latter two categories as "primaries." Because this terminology conflicts violently with Census Bureau usage, the IPUMS avoids this term.

These sample units are largely incompatible with the modern census concept of group quarters, one reason the 1900 sample should probably be redone. The sample does, however, provide sufficient information to determine whether any individual would have been sampled as a group quarters resident in another census year. With care, most common measures can be made compatible with other census years. The design yielded a flat 1-in-760 sample of individuals and families.

1910. The 1910 population census schedules are contained on approximately 1 million census pages of one hundred lines each.

Each reel of the 1910 census was divided into five page segments (or strata), and two randomly chosen lines were designated as sample points in each stratum. Sample units were entered only if they landed on a head or head-equivalent in a regular household, the head or head-equivalent in the primary family of a large household, or an individual unrelated to the head of a large household. The definitions for these units were as follows:

- **Regular household:** census families with a head or head-equivalent and fewer than twenty-one members unrelated to the head.
- **Primary family in large census family:** the head or head-equivalent of a census family and persons related to the head in census families with more than twenty members unrelated to the head.
- **Unrelated individual in large census family:** all individuals unrelated to the head in census families containing twenty-one or more members unrelated to the head.

As in the case of the 1900 sample, I have altered the terminology used in the 1910 documentation.

This procedure yields a representative 1-in-250 sample of households and individuals and, unlike the 1900 design, can be made compatible with later definitions of households and group quarters. In addition to the flat 1-in-250 sample, an oversample of the black population in 1910 is available, and we are presently working on an oversample of the Hispanic population.

1920. The manuscript census of the 1920 population consists of about 1.2 million census pages recorded on 2,076 reels of microfilm. Each census page has one hundred lines, and the information pertaining to each individual appears on a separate line.

We adopted the same sampling scheme in 1920 as in 1850, except that we used family relationships as well as surnames to identify related groups in group quarters.

1940. The population schedules of the 1940 census are preserved on 4,576 microfilm reels. Each census page contains information on forty individuals. Two lines on each page were designated as "sample lines" by the Census Bureau; the individuals falling on those lines—5 percent of the population—were asked a set of supplemental questions that appear at the bottom of the census page.

Two of every five census pages were systematically selected for examination. On each selected census page, one of the two designated sample lines was then randomly selected. Data-entry personnel then counted the size of the sample unit containing the targeted sample line. Units size six or smaller were included in the sample in inverse proportion to their size. Thus, every one-person unit was included in the sample, every second two-person unit, every third three-person unit, and so on. Units with seven or more persons were included with a probability of 1-in-7: every seventh household of size seven or more was selected for the sample.

Sample units for 1940 were defined as follows:

- **Households:** dwelling places with fewer than five persons unrelated to a household head, excluding institutions and transient quarters.
- **Group quarters:** individuals in dwelling places with five or more persons unrelated to a household head, and individual residents of institutions and transient quarters.

This design ensures that each selected sample unit contains one individual who was asked the supplemental sample questions at the bottom of the enumeration form. It yields a flat 1-in-100 sample of persons in units of size seven or less. Persons in units larger than seven are overrepresented in the 1940 PUMS; they must be weighted downward to achieve a representative distribution of household size. Analyses of sample-line individuals who answered supplemental questions must also be weighted. Appropriate weights are included in the sample.

1950. The 1950 census schedules are contained on 6,278 microfilm reels. Each census page contains information on thirty individuals. Every fifth line on the census page was designated as a sample line, and additional questions for the sample-line individuals on each page appear at the bottom of the form. For the last sample-line individual on each page, there was a block of additional supplemental questions. Thus, 20 percent of individuals were asked a basic set of supplemental questions, and 3.33 percent of individuals were asked a full set of supplemental questions.

One-in-eleven pages within enumeration districts was selected randomly. On each selected census page, the sixth sample-line individual (the one with the full set of questions) was selected for inclusion in the sample. Any other members of the sample unit containing the selected individual were also included. Sample units are defined as in the 1940 sample.

As in the 1940 sample, each household in the 1950 sample includes one individual who was asked supplemental questions. The sampling procedure yields a flat 1-in-330 sample of these sample-line individuals. But the sampling procedure is not flat for persons who were not sample-line individuals. The probability of inclusion in the sample is directly proportional to the size of the unit. Thus, when analyzing the entire population of the persons in units with more than one individual, cases must be weighted in inverse proportion to household size. An appropriate weight is included in the sample.

1960. The 1960 census used a machine-readable household form instead of the traditional census schedule. Census information was collected on a separate form for each housing unit. For the first time, the housing questions were included on the same form as the population items. Every fourth enumeration unit received a *long form,* which contained supplemental sample questions that were asked of all members of the unit. Since the public use microdata files are drawn entirely from these long forms, the sample questions are available for all individuals in every unit, instead of a single member of each unit as in 1940 and 1950. Four-fifths of the enumeration units received one version of the long form (the 20 percent questionnaire), and one-fifth received a second version (the 5 percent questionnaire) with slightly different housing questions. The 1-in-100 1960 PUMS is drawn from both questionnaires. Sample units in 1960 were defined the same as the sample units of the 1940 PUMS.

The selection procedure for including cases from the 25 percent sample questionnaires in the public use sample was carried out in three steps. First, the entire census was divided into 33,000 geographic units, called smallest weighting areas (SWA). The population of each SWA was broken into forty-four categories, based on broad age group, sex, race, headship, and homeownership, and for each category a weight was calculated representing the ratio of persons in the full population count to persons in the 25 percent sample. These weights were used in calculating most census tabulations of sample characteristics for small geographic areas.

In the second step of sample selection, the sample weights generated for each SWA were used to select a stratified 5 percent sample from the 25 percent sample. The 25 percent sample of the long forms was divided into thirty-eight strata, based on household size, homeownership, race, and group quarters residence. Within each stratum, the cumulative sum of weights for each household head was calculated, and a case was selected for inclusion in the sample each time the cumulative sum passed a multiple of twenty. This procedure yielded a flat 5 percent sample that was used to produce many of the census publications pertaining to the general population.

Finally, the 1 percent sample was selected from the 5 percent sample, using essentially the same procedure to select every fifth case within each of thirty-eight strata. The strata used in this selection were the same as those used to select the 5 percent sample, except that a slightly different classification of household size is used . The 1 percent 1960 sample is divided into one hundred subsamples, each of which incorporates the same stratification. The three-step elaborate selection scheme for the 1960 Public Use Sample yielded a flat sample with very small standard errors, especially for race and homeownership.

1970. One-in-five housing units in 1970 received a long form containing supplemental sample questions. Sample units were defined the same way as the sample units of the 1940 PUMS. There were two versions of the long form, with different inquiries on both housing and population items; 15 percent of households received one version, and 5 percent received the other. Six independent 1 percent public use samples were produced for 1970, three from the 15 percent questionnaire, and three from the 5 percent questionnaire. Each of the three samples drawn from each questionnaire provides somewhat different geographical information.

The procedures used to select cases for inclusion in the 1970 Public Use Samples were similar to those used in 1960 but were slightly more elaborate. Again, weights were constructed for the SWA as the ratio of persons with selected characteristics in the full population count to persons with the same characteristics in the 15 percent and 5 percent samples. In 1970, these weights were calculated in three stages that controlled for household size, sex of head, presence of own children of head, group quarters residence, headship, race, age, and sex.

To select the six 1 percent samples from the 15 percent sample and the 5 percent sample, the Census Bureau divided the weighted population for each sample into seventy-five strata, based on homeownership, race, sex of head, household size, presence of own children, inmate status, and other residence in group quarters. Within each stratum, the sum of weights for household heads was cumulated. The weights represent the ratio of persons in the full count to persons in each sample; because three 1 percent extracts were required for each sample, a case was selected each time the cumulated total of weights passed a multiple of thirty-three. As in 1960, each sample was divided into one hundred subsamples all of which incorporate the same stratification.

1980. The 1980 census employed a single long-form questionnaire completed by half of housing units in places under twenty-five hundred population and one-sixth of other housing units. Overall, 19.4 percent of housing units were included in the sample. Sample units were defined the same as in 1970, except that the threshold for sampling as group quarters was raised from five or more persons unrelated to the head to ten or more persons unrelated to the head. Three PUMS were produced in 1980: a 5 percent sample and two 1 percent samples. Each of the three samples has differing geographic detail.

The 1980 census used the same procedures as the 1970 census to select long-form sample cases for inclusion in the PUMS, but each step was more elaborate. As in 1970, a three-stage ratio estimation procedure was used to assign weights to sample cases representing the ratio of the full population count to the sample count for persons with particular characteristics in smallest weighting areas. This time, the weights were designed to control for 179 characteristics and combinations of characteristics, including household size, presence of own children, group quarters residence, householder status, detailed race and Spanish origin, age, and sex. The weighted population was divided into 102 strata, including breakdowns by race, Spanish origin, homeownership, sampling rate, and presence of own children. As in 1960 and 1970, cases were selected by cumulating the weights within each stratum, and one hundred stratified subsamples were identified within each of the 1980 PUMS.

1990. The 1990 census used a single long-form questionnaire for sample questions completed by half the persons in governmental places under twenty-five hundred population, one-sixth of persons in other tracts and block numbering areas with fewer than two thousand housing units, and one-eighth of all other areas. Overall, about 1-in-6 housing units completed a long form. Sample units were defined the same as in 1980. Three PUMS files were produced: a 5 percent sample, a 1 percent sample containing somewhat different geographic codes, and a 3 percent sample of the elderly.

The ratio estimation procedure used to assign weights to sample cases in 1990 was virtually identical to the procedure used in 1980. The stratification scheme used to select cases for inclusion in the PUMS, however, continued the trend toward increasing complexity: the number of separate strata was increased from 102 to 1,049, mainly because of additional detail on age and race.

At this point, the 1990 selection procedure broke with the precedent established in the previous three census years. The previous censuses used the weights to extract a flat sample from each stratum, so the final public use samples had equal probabilities of inclusion for all individuals and households. For 1990, the Census Bureau opted instead to produce a weighted PUMS file. Within each state, the Bureau divided the sample questionnaires into an appropriate number of 1 percent samples. For example, if 20 percent of the population of a state completed long forms, the sample questionnaires for that state were divided into twenty subsamples of equal size. Each subsample would then consist of every twentieth case drawn from each stratum. The 5 percent, 1 percent, and 3 percent PUMS files were then selected at random from the 1 percent subsamples for each state. Weights were attached to each case representing the number of individuals in the general population represented by any particular case in the sample; these weights range from 0 to 1,138.

The advantage of the weighted sample design adopted for 1990 is that it provides maximum precision for persons residing in small localities. The disadvantages are great, however. The sample is not only far more cumbersome to use than those previously produced by the Census Bureau, but precision is actually reduced for the general population. For these reasons, we are exploring methods for adding a variable to the 1990 IPUMS file that would allow users to extract an unweighted representative subset of the data.

NOTES

1. These are the modern census terms. The census has historically employed a variety of terms and definitions for distinguishing housing units. Before the modern census concepts of household and group quarters, past censuses distinguished quasi-households, families, and dwellings. There have been subtle variations in the definitions of all these terms from census year to census year; see Appendix. These changes in definitions have modest consequences, and very close approximations of the modern concepts of household and group quarters can be reconstructed for all census years since 1850.

2. For further discussion of this method for estimating design factors, see U.S. Bureau of the Census (1993). All variables were treated as categorical variables, so the design factors reproduced here are actually the weighted average of the factors for each category of each variable. For the most part, I used the "general" IPUMS version of each variable, except that age was grouped in ten-year intervals and top-coded at 80, and occupation, language, and birthplace used only the first two digits of the IPUMS variable. The fifty subsamples were derived from the one hundred subsample replicates provided in the censuses of 1960 onward and in the IPUMS for earlier years; in all cases, selection of subsamples replicated the original sample selection procedures. For 1940 and 1950, I used the self-weighting versions of the samples to estimate design factors. The estimates presented here differ from those provided by the Census Bureau for the 1970 and 1980 PUMS, but they are probably more reliable. The Census Bureau estimates were created before the samples were available and were derived from a simplified analytic model; see U.S. Bureau of the Census (1973, 1983).

PART 2. Order Out of Chaos: The Integrated Public Use Microdata Series

The Comparability of Occupations and the Generation of Income Scores

Matthew Sobek

The Integrated Public Use Microdata Series (IPUMS) is committed to the concept of comparability over time. With respect to occupations, this commitment entailed imposing a standard occupational classification scheme in all census years from 1850 to 1990. The common coding allows researchers to use 1950 occupations and their groupings as a means of locating persons in the occupational/social structure. Without such consistent coding, it is difficult to make temporal comparisons. Despite the information conveyed by such a classification, however, occupation remains an unwieldy variable. Many researchers have found occupational groupings too general and heterogeneous for their purposes, as well as unsuitable for many statistical techniques. An alternative approach involves scaling occupations according to some external criterion in order to turn occupation into a measure of prestige or economic standing. Such measures are a staple of modern social scientific research. Using total income in 1950, we constructed such a measure for the IPUMS. This economic score represents the material rewards accruing to persons in different occupations.

Occupational Coding

Occupation is among the most problematic census variables in terms of comparability. Every Public Use Microdata Sample (PUMS) project coded occupation into the contemporary Census Bureau classification scheme, which changed considerably over time. Additionally, each PUMS prior to 1940 also coded the manuscript occupation responses into more modern classification schemes. The 1850, 1880, 1900, and 1920 PUMS presented occupations in the 1950 system, and the 1910 PUMS used the 1980 scheme. The importance of occupation combined with the variety of classifications presented the IPUMS project with one of its most challenging coding tasks.

The Census Bureau has a history of tinkering with occupational classifications from decade to decade. In some years—1910, 1940, and 1980—the Bureau dramatically reorganized the entire scheme. Table 1 presents some examples of different groupings used by the Bureau. Occupational classifications developed in the nineteenth century were more oriented toward the work setting and economic sector than to a person's specific technical function. Such classifications might reveal that someone worked in the iron-and-steel industry or was a railroad employee, without identifying his specific task or position. By 1910, the number of categories exploded as the scheme tried to incorporate both function and setting (e.g., laborers, marble and stone yards). In 1940, the Census Bureau finally adopted the socioeconomic classification of occupations championed by longtime Census agent and researcher Alba Edwards (1938). Work setting and economic sector were largely relegated to a separate industry variable.

In addition to changes in classification, the universe of persons at risk of having an occupation shifted subtly in 1940. Before then, a person was recorded as having an occupation if he or she was "gainfully employed" in the previous year. This amorphous concept posed particular problems of interpretation with respect to children, women, and seasonal employment (Smuts 1960; Moen 1988; Folbre and Abel 1989). In 1940, application of the labor-force con-

TABLE 1
Select Census Bureau Occupational Groupings, 1880–1990

1880	1910	1950	1990
Agriculture	Agriculture	Professional	Managerial and professional
Professional and personal service	Extraction of minerals	Farmers	Technical, sales, and administrative support
Trade and transportation	Manufacturing and mechanical	Managers, officials, and proprietors	Service
Manufacturing, mechanical, and mining	Transportation	Clerical	Farming, forestry, and fishing
	Trade	Sales	Precision production, craft, and repair
	Professional service	Craftsmen	Operators, fabricators, and laborers
	Domestic and personal Service	Operatives	Military
	Clerical occupations	Service	
		Farm laborers	
		Laborers	

cept defined participation as having worked for pay at any time within a particular reference week. Unpaid family workers had to meet a threshold of a certain number of hours to be classified as part of the labor force.

In order to make a compatible classification for the IPUMS, we had to choose a particular year as a standard. For a number of practical reasons, we picked the 1950 occupational coding scheme. The 1950 census is more or less in the middle of the PUMS series. Three of the four early PUMS were already coded into the 1950 classification. The 1950 system exemplifies the status-hierarchy classification with which social scientists have grown familiar (i.e., professionals, clericals, skilled workers, laborers). A great deal of historical work has either explicitly or implicitly relied upon this scheme. Stephan Thernstrom's (1973) path-breaking mobility study, *The Other Bostonians*, relied upon a reworking of these mid-twentieth-century groupings applied back to nineteenth-century data. We also constructed a separate, compatible industry variable using the 1950 system to provide additional information on work setting and economic sector (e.g., furniture and fixtures, manufacturing).

We retained the unchanged contemporary occupation codes in a separate variable. For researchers not interested in change over time, the contemporary classification may be the most useful. These coding systems lack anachronisms and make distinctions that may have been especially meaningful in their respective times. Furthermore, the intersection of the contemporary historical scheme with the 1950 classifications can yield information not separately contained in either. For instance, the 1950 code may describe the task performed (spinner), while the 1880 code gives the industry (cotton-mill operative). Actually, this additional information applies only to the historical PUMS, from 1850 to 1920. In these datasets, the 1950 codes are not simply recodes from the contemporary system but were assigned independently, directly from the alphabetic manuscript responses.

Our coding into the 1950 system was greatly aided by a number of technical papers published by the U.S. Bureau of the Census (1968, 1972b, 1989). Since 1950, the Bureau has provided data documenting the effects of classification changes. The data give the distribution of persons in a particular 1970 occupation, for example, as they would have been coded under the 1960 system. Using the technical papers, we recoded the occupation based on where a plurality of persons (the largest single group) would have been assigned had the previous decennial census coding system been used. We thus tracked from decade to decade, back to 1950, where a plurality of persons would have been coded. These are the 1950 occupations recorded in the IPUMS. It should be noted that this "tracking" of the plurality assumes the even distribution of persons among the pieces that subsequently get tracked to the next decade (a concern relevant only to 1970, 1980, and 1990).

With only a few anomalies (some residual categories and a few omissions), the technical papers are comprehensive in detailing occupational change. However, the technical papers only cover the period from 1950 to 1980. Fortunately, the 1990 system was identical to the one used in 1980, with very minor exceptions. The change between 1940 and 1950 was also limited, thus requiring very little judgment on our part. The pre-1940 PUMS—with the exception of 1910—were already coded directly into the 1950 system from the original alphabetic responses. This left 1910 as the only truly difficult year. Nineteen-ten was coded into the 1980 system, so we could have used the 1980 codes and then applied the 1980 pluralities backward to 1950. Instead, we simply recoded from the 1910 titles directly into the 1950 system solely on the basis of occupational title. This recoding was feasible because of the extreme detail of the 1910 system. In 1910, the Census Bureau had not yet separated the industrial from occupational classifications. The 1910 scheme thus contained a great deal of information about occupational work setting as well as technical function. The vast majority of coding was straightforward or

was accomplished using the 1950 *Alphabetical Index of Occupations and Industries* (U.S. Bureau of the Census 1950). Little subjective judgment was called for.

In carrying out our recoding, we make no claims as to the continuity in the social setting or technical content of specific occupations. There has obviously been significant change within many occupations. It is up to individual researchers to decide when the 1950 system of classification is appropriate to their use, and what caveats or modifications might be in order. We believe the occupational groupings (professional, craftsmen) within the 1950 system can be used with a high degree of confidence in terms of relative social status and general function. Furthermore, our experience suggests that the specific occupational titles are less subject to meaningful change than the common historical wisdom would suggest (Sobek 1991). We think that the 1950 codes generally work well as a means of locating individuals in the occupational structure as far back as the mid nineteenth century.

Occupational Income Scoring

Despite the utility of the common 1950 classification, some scholars desire a numeric indicator of occupational status suitable for advanced statistical techniques. For the IPUMS, we constructed an income score based on the relative economic standing of occupations in 1950. Median incomes for each occupation were calculated from the data published by the U.S. Census Bureau (1956) *Special Report* on occupational characteristics. A 3.33 percent sample of the population provided the data for the report. We combined the median figures for men and women, which were presented separately in the published data. The 1950 occupational income we use is the weighted mean of the two median figures in hundreds of 1950 dollars. The data are based on the total income (not merely wage and salary) of all persons in the given occupation in 1950. Although in all cases we used 1950 income data, the score was calculated differently for the pre-1950 and post-1950 periods. Understanding this difference is critical for correctly interpreting the score.

The IPUMS occupational income score is intended to account for some of the effects of classification changes over time after 1950. As described above, we used Census Bureau technical documentation to recode post-1950 occupations into 1950 occupations by tracking where the largest subgroup of the occupation would have been coded under each preceding system. In contrast to the occupation recode, however, the income score retains all the various components of the occupation as it gets tracked from census to census, back to 1950. The income score is a weighted average of the incomes of these occupational components. Therefore, the income score describes the 1950 median income of the weighted average of the constituent parts of each occupation (from 1960 to 1990), as coded into the 1950 system.

Table 2 demonstrates the process of tracking occupations and weighting the scores by the economic standing of occupations in 1950. Column A shows the three occupations into which 1980 patternmakers would have been coded under the 1970 system. Column B breaks each 1970 occupation into its 1960 components. Column C tracks 1960 occupations into 1950, providing a list of all the 1950 occupations into which 1980 patternmakers would have been classified. Column D gives the final distribution of 1980 patternmakers among 1950 occupations after tracking the pieces over all intervening years. The final score for 1980 patternmakers is 35 ($3,500 in 1950). Without the weighting procedure, the plurality-based method would have resulted in the 1950 patternmaker score of 34. The inclusion of persons in 1980 who in previous years would have been identified as toolmakers and designers raised the score by one point after weighting. This procedure was carried out separately for every occupation in each census from 1960 to 1990.

The pre-1950 income scores were calculated differently. For the years prior to 1950, we did not have the benefit of documentation permitting the aforementioned weighting procedure. Consequently, the 1950 occupation codes are the sole basis of income score assignment. Fortunately, the 1950 classification system is more amenable to direct recoding from the pre-1950 period than it is from the post-1950 period, when the coding scheme changed markedly.

The question of change over time in the relative incomes of occupations should elicit caution among users. For the post-1950 period, weighting controls for classification, but not for change, in occupational income over time. The applicability of the 1950 scores to 1910 and before is open to question, given the length of time and the lack of individual income data. This issue might be resolved by empirical investigation using other sources.

Although the income score is derived from individual-level data, it should not be interpreted as actual personal income. We conceive of the score as a method of scaling occupations—essentially a way of turning occupation into a continuous measure. An occupation with a high score is a well-rewarded and probably high-status occupation. *The measure is an economic score, not a socioeconomic one.* Depending on one's perspective, some aspects of social status may be better reflected in the Census Bureau's grouping of occupations in 1950 than in the IPUMS income score. The divide between manual and nonmanual work or the identification of craftsmen (skilled labor) are among the distinctions better made through occupational titles and groups.

The IPUMS income score is an objective measure of status. There is no prestige dimension to the income score, in contrast to the widely used Duncan Socioeconomic Index (SEI), which is also pegged to the 1950 coding system (Duncan 1961). The SEI is an indicator of occupational status combining income and educational attainment to predict the results from a 1947 survey on the "general standing" of

TABLE 2
Derivation of the Income Score for the Patternmaker Occupation in 1980

A: 1980 patternmakers broken down into 1970 occupational components			B: 1970 component occupations broken down into 1960 occupational components			C: 1960 component occupations broken down into 1950 occupational components		D (A*B*C) Proportion of 1980 patternmakers in each 1950 occupational component	E Income Scores of 1950 occupational components	F (D*E) Income score weight by component
%	Occ.		%	Occ.		%	Occ.	%		
17.57	Designers	>	93.00	Designers	>	100.00	Designers	16.34	38	6.21
		>	2.20	Professionals[a]	>	90.10	Professionals[a]	.35	33	.12
					>	8.09	Attendants, service	.03	13	.00
					>	1.39	Architects	.01	54	.01
					>	.41	Agents[a]	.00	36	.00
		>	4.79	Operatives[a]	>	99.78	Operatives[a]	.84	23	.19
					>	.16	Truck drivers	.00	25	.00
					>	.02	Excavators	.00	31	.00
					>	.02	Asbestos workers	.00	32	.00
					>	.02	Craftsmen[a]	.00	32	.00
48.12	Patternmakers	>	96.76	Patternmakers	>	100.00	Patternmakers	46.56	34	15.83
		>	3.23	Operatives[a]	>	99.78	Operatives[a]	1.55	23	.36
					>	.16	Truck drivers	.00	25	.00
					>	.02	Excavators	.00	31	.00
					>	.02	Asbestos workers	.00	32	.00
					>	.02	Craftsmen[a]	.00	32	.00
34.31	Tool- and die-makers	>	95.13	Toolmakers	>	100.00	Toolmakers	32.64	37	12.08
		>	4.87	Machinists	>	100.00	Machinists	1.67	32	.53
									Final weighted occupational score (sum of F):	35.33

Notes:
Column A: The occupational distribution of 1980 patternmakers as they would have been classified in 1970.
Column B: The occupational distribution of each 1970 component occupation as it would have been classified in 1960.
Column C: The occupational distribution of each 1960 component occupation as it would have been classified in 1950.
Column D: The proportion of 1980 patternmakers as they would have been classified in 1950 (column A * column B * column C).
Column E: The median income (in hundreds of dollars) of the 1950 component occupations in column D.
Column F: The amount of 1980 income score contributed by each 1950 component occupation (column D * column E).
[a]Not elsewhere classified.

occupations. There are inherent problems in relying on such subjective judgments of occupational standing, not the least of which is the lack of comparable survey data for earlier periods that might suggest the degree of change over time. Scholars have engaged in a great deal of debate over the years concerning the relative merits of objective and subjective determinants of occupational status (Hodge, Siegel, Rossi 1964; Penn 1975; Featherman and Hauser 1976; Treiman 1976; Horan 1978; Hauser 1982; Nam and Powers 1983). One conclusion that can be drawn from the literature is that most of what a prestige measurement captures is, in any case, economic standing. An objective score like income presents fewer interpretive problems. Even if one is theoretically inclined to prefer prestige measures, however, one would be hard pressed to find the necessary data in the historical record. We chose a score consistent with the essentially objective nature of the remainder of the census data.

There is much ground for theoretical debate concerning the use or misuse of occupational status measures, but the above qualifications and cautions should not give the wrong impression. The income score works, producing plausible results in every study we have undertaken using it. It is among the most powerful explanatory variables available in studies on topics as varied as homeownership, school attendance, and family structure (Ryden 1994; Ruggles forthcoming; Kallgren forthcoming). The variable is robust across decades, whatever changes may have occurred in the relative standing of specific occupations. Some may object to such a blatantly empiricist justification, but the variable's strength is undeniable. The real grounds for debate are the measure's degree of imprecision and how much the occupational economic hierarchy changes over time. Future research may answer these questions, but for the present we know that the income score works very well. The only competitors for ordering occupations are the census classifications, intuitive judgments of individual scholars, or prestige measures like the SEI. If economic status is desired, the IPUMS income score is a better indicator than any of these.

PART 2. Order Out of Chaos: The Integrated Public Use Microdata Series

Family Interrelationships

Steven Ruggles

The Public Use Microdata Samples (PUMS) of the census are simultaneously samples of households and of individuals, and within households the interrelationships among individuals are known. This hierarchical structure is one of the greatest strengths of the census files. By combining the characteristics of several individuals within a household, researchers can create a broad range of new variables about family and household composition and the characteristics of family members. For example, we can analyze fertility by attaching the ages of all own children to their maternal records, and we can address the family economy by simultaneously measuring the age, sex, and occupation of all family members.

Each of the original PUMS provides constructed variables describing household and family composition and family interrelationships. In general, however, these variables are incompatible with one another. Moreover, the household and family variables provided with the original PUMS tend to be inflexible and awkward to use, and they are sometimes inaccurate. One of the goals of the Integrated Public Use Microdata Series (IPUMS) has been to develop a consistent, versatile, and reliable set of tools to make it easy for researchers to construct family variables using standard statistical packages. Unlike many of the original samples, the IPUMS does not include detailed classifications of household and family composition; instead, we supply the basic building blocks for researchers to create their own classifications.

Family Interrelationship Variables: SPLOC, MOMLOC, and POPLOC

Each of the PUMS from 1880 onward includes a variable on the relationship of each household member to the head of household, and a simplified version of this variable can be made fully consistent across all census years from 1880 to 1990 (see article by Ruggles, Hacker, and Sobek on pages 33–39). This variable—called RELATE in the IPUMS—provides the basic measure of family relationships, but it is not sufficient to identify all family relationships and it is often inconvenient as a tool for constructing new family variables. Consider the household in table 1. The relationship variable is sufficient to establish that the two daughters are both children of the household head, but to identify the other family interrelationships we must look to the daughters' other characteristics. We can infer that the son-in-law is married to the second daughter rather than the first one because they share the same surname and are both listed as married; for analogous reasons, we know that the grandchild is probably the child of the second daughter listed. It is also safe to assume that the two boarders are married to one another, because they are both married, they share the same surname, they are both adults and close to the same age, and they are listed adjacently.

To allow users to identify relationships among spouses, parents, and children without forcing them to use multiple variables and complicated logic, the IPUMS includes a set of pointers called SPLOC, MOMLOC, and POPLOC.[1] These pointers identify the location within the household of each individual's own spouse, mother, and father, respectively. Table 2 illustrates these variables. PERNUM is the sequence number of each individual within the household. SPLOC shows the sequence number of each individual's

TABLE 1
Example of Family Relationships

Surname	Relationship	Age	Sex	Marital status
MULCAHY	HEAD	61	F	W
MULCAHY	DAUGHTER	32	F	S
RYDEN	SON-IN-LAW	32	M	M
RYDEN	DAUGHTER	27	F	M
RYDEN	GNDCHILD	4	M	S
SALERNO	BOARDER	26	M	M
SALERNO	BOARDER	22	F	M

TABLE 2
Example of Family Relationships

Surname	Relationship	PERNUM	SPLOC	MOMLOC	POPLOC
MULCAHY	HEAD	01	00	00	00
MULCAHY	DAUGHTER	02	00	01	00
RYDEN	SON-IN-LAW	03	04	00	00
RYDEN	DAUGHTER	04	03	01	00
RYDEN	GNDCHILD	05	00	04	03
SALERNO	BOARDER	06	07	00	00
SALERNO	BOARDER	07	06	00	00

own spouse; for example, since the son-in-law is married to the second daughter who is in the fourth position, his SPLOC is 04. Persons without a spouse are assigned a SPLOC of 00, the usual IPUMS code for not applicable. MOMLOC and POPLOC show the sequence numbers of own mothers and own fathers; for example, the mother and father of the grandchild are in positions 04 and 03, respectively.

SPLOC, MOMLOC, and POPLOC can be used to identify conjugal units, to attach characteristics of spouses or parents, to develop specialized own-child measures, or to serve as building blocks for more elaborate measures of family composition. In most cases, users will be able to manipulate these variables to construct their own measures within a statistical package and will not be forced to resort to higher-level programming.[2]

Most family classification schemes are built up from information on the presence of immediate kin. The basic Census Bureau classifications focus on the presence of spouses and children of the household head; the Laslett (1972) scheme widely used by historians is based on a count of "conjugal family units" consisting of parents and children or married couples. SPLOC, MOMLOC, and POPLOC make it relatively simple to construct such classifications.

Family historians are increasingly moving from household-level schemes of family classification toward individual-level measures of family structure. For example, instead of measuring the proportion of households headed by a single female parent, we might assess the proportion of women who were single parents or the proportion of children residing with mothers only. Such individual-level analyses offer a variety of advantages that have been detailed elsewhere (King and Preston 1990; Ruggles 1987, 1994a, 1994b). The individual-level IPUMS pointer variables are especially well suited to creation of these kinds of measures.

Additional Constructed Family Variables

In addition to SPLOC, MOMLOC, and POPLOC, the IPUMS provides a variety of other constructed variables to aid researchers in creating new family variables. These are described in table 3. Four of the constructed variables apply to entire households. NFAMS is a count of the number of families present in the household. For this purpose a family is defined as any group of persons with identifiable relationships by blood or marriage. A single individual residing without any relatives is considered a separate family. Thus, a household consisting of an elderly widow residing with a servant would count as two families, and a large extended family with multiple generations but no boarders, lodgers, or servants would count as a single family. COUPLES, MOTHERS, and FATHERS are based on counts of SPLOC, MOMLOC, and POPLOC.

TABLE 3
Variables on Family Interrelationships

Variable	Description
	Household record
NFAMS	Number of families in household
COUPLES	Number of married couples present in household
MOTHERS	Number of women with own child present in household
FATHERS	Number of men with own child present in household
	Person record
PERNUM	Sequence number of person within household
RELATE	Relationship of person to household head
FAMSIZE	Number of household members related to person
FAMUNIT	Family unit membership
SPLOC	Location of own spouse within household
MOMLOC	Location of own mother within household
POPLOC	Location of own father within household
NCHILD	Number of own children in household
NCHLT5	Number of own children under age 5 in household
ELDCH	Age of eldest own child in household
YNGCH	Age of youngest own child in household
NSIBS	Number of own siblings in household

The additional individual-level constructed variables on family and household relationships listed in table 3 are illustrated by example in table 4. FAMSIZE and FAMUNIT

TABLE 4
Additional Constructed Family Variables

Surname	Relationship	FAMSIZE	FAMUNIT	NCHILD	NCHLT5	ELDCH	YNGCH	NSIBS
MULCAHY	HEAD	05	1	2	0	32	27	0
MULCAHY	DAUGHTER	05	1	0	0	99	99	1
RYDEN	SON-IN-LAW	05	1	1	1	04	04	0
RYDEN	DAUGHTER	05	1	1	1	04	04	1
RYDEN	GNDCHILD	05	1	0	0	99	99	0
SALERNO	BOARDER	02	2	0	0	99	99	0
SALERNO	BOARDER	02	2	0	0	99	99	0

use the same definition of family employed for NFAMS. FAMSIZE is useful for creating a variety of family measures. For example, to determine if a family contains extended kin beyond spouse and children, one can subtract the size of the immediate family (spouse and children) from FAMSIZE; if the result is greater than one, there are other kin present. More complex measures of extended family configurations can be constructed using FAMUNIT, which in combination with SERIAL provides a unique identifier for each related group in the census. The IPUMS also includes the four most commonly requested measures of own children, all of which are derived from MOMLOC and POPLOC.[3] Finally, there is a basic variable on number of own siblings. Siblings are defined as persons who share a common parent or who have family relationship codes that imply a sibling relationship.

Creation of MOMLOC and POPLOC

Assigning links between parents and their children is usually straightforward. In about 97 percent of cases the census information on family relationships is sufficient to establish parent-child links. For example, if an individual is listed as a child of the household head, his or her parents should always be listed as the household head or wife of head; there is little ambiguity because each household has one head and no more than one wife. Similarly, the parents of persons listed as the household head or a sibling of the head are always listed as mother or father of the head, and each household contains no more than one person listed as mother and no more than one listed as father.[4] Parentage is almost as clear-cut for persons listed as wife or sibling-in-law, since households ordinarily do not include multiple mothers-in-law or fathers-in-law.

For persons who have family relationships other than head, wife, child, sibling, or sibling-in-law, the relationship information does not identify parental relationships with as much precision. For example, we know that the parent of a person listed as grandchild of the head should be listed as a child or a child-in-law, but because a family may contain multiple persons listed as child or child-in-law, the relationships do not unambiguously identify parentage. Even if there is only one child present, there is still room for error, since a grandchild could be the offspring of an absent child. In some cases—such as secondary families consisting of boarders—the relationship codes may provide no information for linking parents and children.

Whenever the family relationship codes are unclear, we must turn to other information to identify parent-child relationships. In every census year from 1880 to 1990, the census contains three additional pieces of information that can be used to clarify ambiguities: age, marital status, and the order in which individuals are listed in the census.[5] Thus, for example, if a household contains a widowed daughter followed immediately by a grandchild who is twenty years younger than the daughter, we may reasonably infer a maternal relationship even if other daughters are present. Each census year also includes other information that can be used to distinguish parental relationships, but the availability of this information is irregular. For example, in the census years 1880, 1910, 1920, 1940, and 1950, we can identify persons who share the same surname. For 1970, 1980, and 1990, on the other hand, we can identify the number of children ever born to every adult woman.

Our procedure for linking parents and children attempted to reconcile two competing goals. The first goal was to create fully compatible links by using only information available across all census years; the second goal was to provide the most accurate possible links in each census year by using all available information. We developed a set of logical rules for establishing parental links that represent a compromise between these conflicting goals. These rules are described in detail in the Appendix.

Giving priority to compatibility, we therefore begin by establishing all parental relationships that can be plausibly identified using only information on relationship, age, sex, marital status, and sequence in the household listing. This is carried out by means of three logical rules. We then use three additional rules to add parental relationships that can be identified only by turning to other information, such as surname similarity, children-ever-born, or children surviving. We worked out the details of all rules through a process

of experimentation, by comparing the results of the rules with our own judgments about a collection of the most problematic households drawn from several census years. The variables MOMRULE and POPRULE identify which particular logical rule was used to establish a parental relationship in any given case. For analyses comparing multiple census years, users can ensure full compatibility by using only those links that were established under rules 1 through 3.

In practice, the additional information available in particular census years doesn't make a great deal of difference. For the censuses of 1880 through 1960, 99.5 percent or more of parental links were established by means of the first three logical rules.[6] In recent census years, the percentage of cases requiring additional information has risen, because marital status has become less of a determinant of parenthood; by 1990, only 97.9 percent of maternal links were established by means of the first three rules.

Identifying Stepparents

The logical rules used to create MOMLOC and POPLOC link parents to adopted children and stepchildren as well as biological children. This may be appropriate for the study of topics such as the family economy, but for some topics—such as fertility analysis—adopted children and stepchildren should be eliminated whenever possible. Users who wish to limit their analysis to biological children can use the variables STEPMOM and STEPPOP to identify probable biological relationships. The values for these variables are defined in table 5.

Where more than one value for STEPMOM or STEPPOP was valid, the lower value was assigned. To analyze biological children one can eliminate links with a value of greater than zero on STEPMOM or STEPPOP. When comparing successive census years, one should use only values 1 and 2 of STEPMOM and STEPPOP, since they are the only ones consistently available.

Table 6 shows the distribution of STEPMOM across all census years. With the exception of the 1900 and 1910 census years, 2 percent or less of children can be identified as stepchildren or adopted children. The frequency of identifiable stepchildren is somewhat higher in 1900 and 1910, which is not surprising since those census years provide more relevant information than any others. In particular, they are the only years that indicate the number of surviving children for each woman.

The true percentage of stepchildren and adopted children is no doubt higher than is indicated in table 6 in all census years. Because we cannot identify all biological children, own-child fertility estimates derived from the census will be biased slightly. In particular, we would expect that estimates of mothers' ages at childbirth may be a bit low, because second and third wives are on average younger than first wives.

Creation of SPLOC

Spousal links are much easier than parental links. Most households have only one married couple; when more than one married couple is present, proximity is a reliable indicator of who goes with whom. In all census years, married couples are listed adjacently in about 99 percent of cases, and the few exceptions can almost all be resolved through relationship codes.

The spousal links were carried out by means of seven rules described in the Appendix. These rules use only infor-

TABLE 5
Stepparent Flags

Flag	Description
0	Probable biological parent.
1	Age difference between parent and child improbable (outside the range 15–49 for women and 15–64 for men).
2	The link was only established because the parent was married to another parent.[a]
3	Detailed relationship codes explicitly specify a stepparent relationship—information not available in 1960, 1970, or 1980.
4	Mother has zero children-ever-born (or surviving children, in 1900–1910). This flag can be constructed for women in 1900, 1910, and 1940 onward.
5	Detailed relationship codes specify that child is adopted—information only available for 1880, 1900, 1910, and 1920.
6	Child was born before marriage of parent (current marriage in 1900 and 1910, first marriage in 1960 and 1970), and there is a mismatch between parental birthplace on child's record and birthplace of parent. This flag is available in 1900, 1910, 1960, and the 1970 5 percent samples.
7	Number of children present exceeds number-ever-born (or number surviving, in 1900 and 1910), and child was born before marriage of mother (current marriage in 1900, 1910, and 1950; first marriage in other years). Available for women in all years except 1880, 1920, 1990, and the 1970 15 percent samples.

[a]See Appendix. The frequency of value 2 on STEPMOM is lower than the frequency of rule 7 on MOMRULE only because most mothers assigned under rule 7 have an improbable age difference and are therefore assigned a STEPMOM of 1.

TABLE 6
Percentage Distribution of Stepmother Flags (STEPMOM) by Census Year

Flag	1880	1900	1910	1940	1950	1960	1970	1980	1990
0	98.0	96.9	97.3	98.3	98.6	98.8	98.0	98.0	98.5
1	1.8	1.3	1.1	1.4	1.0	1.0	1.3	0.9	0.9
2	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0
3	0.0	0.0	0.0	0.1	0.0	—	—	—	0.6
4	—	0.4	0.4	0.2	0.1	0.0	0.3	0.2	0.0
5	0.2	0.2	0.2	—	—	—	—	—	—
6	—	0.6	0.4	—	—	0.0	0.0	—	—
7	—	0.6	0.6	0.0	0.2	0.2	0.2	0.2	—
N	251,058	47,982	167,614	564,510	925,103	728,001	803,956	785,973	810,633

mation that is available in all census years and are therefore fully compatible. Even though the spousal rules ignore much relevant information available in particular census years—such as surname, marriage duration, and age at first marriage—we nevertheless consider them to be much more reliable than the rules governing parental links.

Comparison of IPUMS and 1910 PUMS Linking Procedures

The only previous dataset to incorporate family relationship variables similar to MOMLOC and SPLOC is the 1910 PUMS (Strong et al. 1989). The creators of the 1910 sample adopted a much more elaborate procedure for creating links between mothers and children and between husbands and wives. In the most obvious cases—those with completely unambiguous relationship codes and no other hint of ambiguity—the 1910 PUMS relied on logical rules to assign linkages. In all other cases, however, the 1910 PUMS turned to a complicated point system based on probabilities. Each characteristic that could be used to identify potential mother-child or husband-wife links—such as similar surnames, relationship codes, age differences, and so on—was assigned a point value based on its power to predict "correct" links in a small hand-linked subset of the data. The sum of these points was then calculated for all potential links in the sample. If the sum exceeded a prespecified minimum, the link was accepted, and if it fell below a prespecified minimum, the link was rejected. When the sum of weights fell in the gray zone, links were carried out by hand, by reexamining the case on the original microfilm.

We experimented extensively with similar probability-based point systems for assigning links, but we found them unsatisfactory. The importance of any particular characteristic depends on its context. For example, surnames assume great significance when the relationship codes are ambiguous, but they should otherwise be ignored. A simple additive point system proved incapable of such distinctions.

The 1910 procedure ran into similar difficulties. Despite the complexity of the probability-based linking system, it was sufficient to identify only the most straightforward links. More than 1-in-5 of individuals in the sample—some seventy-five thousand cases—fell into the gray zone and had to be reexamined by hand. If we had adopted a similar procedure for the IPUMS, it would have meant looking up about 10 million cases individually, which would have multiplied the cost of the IPUMS manyfold.

The logical rules described in the Appendix produce results very similar to those obtained by the 1910 project at a fraction of the cost. Table 7 compares the IPUMS links to the 1910 links.[7] The maternal links obtained through each method differed in 0.66 percent of cases. When the two methods differed, we examined each case and found that in many cases the 1910 links were clearly correct. In most cases, however, the census listings are truly ambiguous, and the links are a matter of guesswork. The spousal links are more clear-cut: the IPUMS and the 1910 procedures produce identical results in over 99.9 percent of cases, even though the IPUMS method ignores all variables that are not available for the entire period from 1880 to 1990.

TABLE 7
Comparison of Maternal and Spousal Links, IPUMS Method, and 1910 Method

	Maternal links (%)	Spousal links (%)
Same results by both methods	99.33	99.92
No link by either method	55.30	64.45
Identical links by both methods	44.03	35.47
Different results by each method	0.66	0.08
Linked by 1910 method only	0.33	0.03
Linked by IPUMS method only	0.30	0.03
Different links by each method	0.03	0.01
Total	100.00	100.00
Number of cases	366,239	188,531

Family Interrelationships before 1880

The censuses of 1850, 1860, and 1870 did not inquire about either relationship to head of household or marital status, the two most important variables for identifying parental and spousal relationships. We are therefore developing an additional set of logical rules for inferring family relationships in those census years, using only information on surname, age, sex, and position in the household listing. A preliminary version of these rules is described in the Appendix. These rules can be tested against the 1880 and 1910 census years so that the reliability of the inferred relationships can be evaluated. In most cases, the inference procedure works well; overall, our preliminary logical inference rules correctly identify 99.4 percent of the spouse relationships and 96.5 percent of the parent-child relationships that can be identified using information on relationship and marital status in 1880.

The logical rules for inferring family relationships are not effective for parents and spouses with different surnames. In general, therefore, we cannot identify parental relationships of ever-married women. For example, the relationships of coresident married daughters or parents-in-law cannot be distinguished without explicit information on family relationships. Thus, the inferred relationships for the period before 1880 exclude some kinds of relationships. To allow consistent comparisons of family interrelationships for the period before 1880 with the period from 1880 onward, we plan to include information on inferred interrelationships even for those years in which more explicit information is available.

APPENDIX

Logical Rules for Inferring MOMLOC, POPLOC, and SPLOC

In the great majority of cases, parents and children can be linked together unambiguously by using a simple set of rules applied to five pieces of information available in every census year from 1880 through 1990: general relationship, age, sex, marital status, and sequence in the household. In a few instances, however, it is necessary to use additional information available only for a subset of census years. The IPUMS linking procedure is designed to allow users to use only links based on information available across all census years, or to use extra information available in a particular census year to make the additional links.

Parental Links

Parental links were established through seven basic rules. If a link could be established through more than one rule, the lower-numbered rule was used. The rule used in any particular case is identified in the variables MOMRULE and POPRULE.

Rule 1: Unambiguous relationships

a) If the relationship of an individual to the household head is son, daughter, or child, then establish parental links to persons listed as head or wife, or

b) if the relationship of an individual to the household head is head, brother, or sister, then establish parental links to persons listed as mother or father, or

c) if the relationship of an individual to the household head is wife, brother-in-law, or sister-in-law, then establish parental links to persons listed as mother-in-law or father-in-law.

Rule 2. Grandchildren

If the relationship of individual to household head is listed as grandson, granddaughter, or grandchild, then establish parental link to the most proximate ever-married child and/or child-in-law with a plausible age difference. Plausible age differences are defined as twelve to fifty-four years for women, and fifteen to seventy-four years for men. If there is more than one eligible parent, choose the most proximate.

Rule 3. All other relatives and nonrelatives (using household position)

Link relatives and nonrelatives not mentioned above to any preceding ever-married person with a plausible age difference as defined in rule 2, as long as there are no intervening persons other than children or spouses of the potential parent. Links between relatives and nonrelatives are prohibited.

Rule 4. All other relatives and nonrelatives (using surname)

Same as rule 3, except that surname similarity is substituted for the requirement that there are no intervening persons between the parent and child. If more than one eligible parent is found, the most proximate is linked. This rule can only be applied in years with surname codes: 1880, 1910, 1920, 1940, and 1950.

Rule 5. Grandchildren (using children born or surviving)

Same as rule 2, except evidence on children-ever-born (or children surviving in 1900 and 1910) is substituted for marital status of parent. This rule does not apply in 1880, 1920, 1940, or 1950.

Rule 6. All other relatives and nonrelatives (using children born or surviving)

Same as rule 3, except evidence on children-ever-born (or children surviving in 1900 and 1910) is substituted for marital status of parent. Again, this rule does not apply in 1880, 1920, 1940, or 1950.

Rule 7. Spouse of linked parent

If one parent is linked and has a spouse present, that spouse is linked as a stepparent.

Users who want to limit their analysis to links that could be recognized in all census years can simply ignore links based on rules 4 through 7. Table A-1 gives the distribution of rules used in each census year. In each year, over 95 percent of links were established on the basis of rule 1. For the period before 1980, over 99 percent of cases were linked on the basis of rules 1, 2, or 3, which are fully compatible across census years. With the increase of births to never-married women in recent census years, however, rules 5 and 6 have become increasingly important, since they substitute information on children-ever-born for information on marital status.

We performed two basic checks for inconsistency of the family links. First, if two parents were linked but they were not married to each other, we unlinked the father. Second, if both partners in a married couple were linked to the same parent, we chose the best parental link based on detailed relationship code, surname, and proximity within the household.

Spousal Links

Spousal links were made based on the following five rules, identified in the variable SPRULE:

1. Link married women to previous adjacent married males with an appropriate relationship. Appropriate relationships are defined as follows:

Relationship	Spouse's Relationship
Head	Spouse
Child	Child-in-law
Parent	Parent
Parent-in-law	Parent-in-law
Sibling	Sibling-in-law

2. Link married women to following adjacent married males with the appropriate relationship.

3. Link married women to nonadjacent married males of appropriate relationship, provided both are over age 15, the husband is no more than twenty-five years older than the wife, and the wife is no more than ten years older than the husband.

4. Link married women with a relationship not specified on the appropriate relationship list to previous adjacent married men with appropriate

TABLE A-1
Percentage Distribution of Maternal Linking Rules (MOMRULE) by Census Year

Flag	1880	1900	1910	1940	1950	1960	1970	1980	1990
1	97.0	97.2	97.3	95.5	93.9	97.0	97.0	96.5	95.1
2	1.6	1.7	1.7	3.1	4.6	2.2	1.5	1.4	1.9
3	1.0	0.6	0.9	1.0	1.2	0.5	0.5	0.6	0.9
4	0.2	—	0.0	0.1	0.0	—	—	—	—
5	—	0.2	0.0	—	—	0.0	0.3	0.9	1.5
6	—	0.0	0.0	—	—	0.0	0.0	0.3	0.3
7	0.2	0.2	0.2	0.2	0.2	0.5	0.8	0.3	0.3
N	251,058	47,982	167,614	564,510	925,103	728,001	803,956	785,973	810,633

ages as defined in rule 3. Ignore relationship, but do not marry an unrelated person to a relative.

5. Same as rule 4, but link subsequent adjacent husbands.

Inferred Links

We have not yet finalized the rules for inferring links to spouses and parents when information on family relationships and marital status is not available, as in the case of the 1850 sample. We are presently working with the following rules (users should check the final IPUMS documentation to obtain the latest version):

The inferred spousal links, like the regular spousal links, are simple and reliable. Potential spouses must be of the opposite sex and share the same surname. Wives must be at least 16 years old and no more than seventeen years older than their husbands; husbands must be at least 18 years old and no more than twenty-eight years older than their wives. First, each eligible male in a household is examined and linked to any eligible adjacent subsequent female; then, each eligible female in the household is examined and linked to any eligible adjacent subsequent male. If any nonadjacent eligible couples remain after these first two passes, they are linked provided the wife is no more than five years older than the husband and the husband is no more than fifteen years older than the wife. If there are more than one potential nonadjacent spouse, the most proximate is chosen.

Inferred parental links are a little more complicated. Potential parents must share the same surname as their children, except for adult daughters surrounded by individuals with the parental surname. Mothers must be between fifteen and forty-nine years older than their children, and fathers must be between sixteen and sixty-four years older. If a child is linked to more than one potential parent of a given sex, priority is given to parents who are married to another parent. Otherwise, priority is given to the first listed potential parent, unless there is a subsequent parental group immediately preceding the child or siblings of the child.

The inferred links are subject to several consistency checks. If two parents are linked and they are not married to one another, the link is severed to the parent listed last. If a married couple shares the same parent, then the spousal link is severed unless the couple has an own-child present, in which case the parental link of the female partner is severed. If two individuals are linked both as parent and child and as husband and wife, the spousal link takes precedence unless the individual listed after the potential child shares the same parent and is within four years of age, in which case the parental link takes precedence. Finally, if a woman is linked to a spouse and to a parent who shares the same surname, the parental link is severed.

NOTES

1. I originally developed this system of expressing family interrelationships in 1980 during my dissertation research, subsequently published as Ruggles (1987).

2. For example, users frequently need to attach the characteristics of immediate family members. The following SPSS-X command file uses SPLOC to attach spouse's occupation to the record of each married person. SERIAL is the IPUMS variable for household serial number, which is a unique identifier for each household. First we obtain an active file with serial number (SERIAL), occupation (OCCUP), and spouse location (SPLOC). SPLOC is renamed as PERNUM, and OCCUP is renamed as spouse's occupation (SPOCC). We sort the file by SERIAL and PERNUM, and then match it back to the original file. Because the PERNUM we are matching was originally SPLOC, we are actually matching spousal occupations.

```
GET FILE='IPUMS.SYS' /KEEP SERIAL SPLOC OCCUP
    /RENAME (PERNUM=SPLOC)(SPOCC=OCCUP)
SORT CASES BY SERIAL, PERNUM
MATCH FILES TABLE=* /FILE='IPUMS.SYS' /BY SERIAL,
    PERNUM
SAVE OUTFILE='IPUMS2.SYS'
FINISH
```

It is virtually as easy to use MOMLOC and POPLOC to attach characteristics of own children. The following SPSS-X command file uses similar logic together with the AGGREGATE command to count the number of own children under the age of 10 for each woman.

```
GET FILE='IPUMS2.SYS' /KEEP SERIAL MOMLOC AGE
    /RENAME (PERNUM=MOMLOC)
SELECT IF (AGE LT 10 AND PERNUM GT 0)
SORT CASES BY SERIAL, PERNUM
AGGREGATE OUTFILE=* /BREAK SERIAL PERNUM
    /CHLT10=N
MATCH FILES TABLE=* /FILE='IPUMS.SYS' /BY SERIAL,
    PERNUM
IF (MISSING(CHLT10)) CHLT10=0
SAVE OUTFILE='IPUMS3.SYS'
FINISH
```

3. NCHILD, ELDCH, and YNGCH are based on all own children; NCHLT5 is a count of own children under 5, excluding identifiable stepchildren and adopted children (see also table 5 and note 7). ELDCH and YNGCH receive a value of 99 if no own children are present, because the usual IPUMS code for not applicable (0) is also a valid age.

4. Households in the PUMS files occasionally include more than one head, wife, mother of head, or father of head, usually because of enumerator or data-entry error. Such cases are corrected in the IPUMS version of the data by means of a consistency checking program prior to assigning parentage. We encountered true polygamous marriages very rarely; where these could be identified, we assigned the wives detailed relationship codes of PG wife.

5. Before 1960, census enumeration instructions specified the sequence in which various relatives and nonrelatives should be listed. With the introduction of self-enumeration in recent census years, the instructions have become less explicit; nevertheless, respondents still have a strong tendency to list children immediately following their parents and to list married couples adjacently.

6. Even a small percentage of missed parental links can have significant consequences for some kinds of analyses, such as the study of young children residing without parents.

7. Since the 1910 variable excluded stepchildren and adopted children, the comparison is based on links with a value of 0 for STEPMOM.

PART 3. From Microfilm to Microdata: Creation of the Public Use Census Files for 1850, 1880, and 1920

Software Development

Todd Gardner

The building blocks for the Integrated Public Use Microdata Series (IPUMS) are national datasets for eleven census years created by the Census Bureau and by individual scholars. Here at the University of Minnesota's Social History Research Laboratory we are responsible for three of those datasets—the 1850, 1880, and 1920 samples. The first step in the creation of each microdata project was software development. Each individual Public Use Microdata Sample (PUMS) required software for data entry, post-entry data consistency checking, verification, and final data recoding. We first began work on the 1880 project and developed all the software on microcomputers. As the volume of data with which we were working increased, however, the performance of the microcomputers became unacceptable for many of these tasks, primarily because of their limited memory. We therefore shifted post-entry data processing to a UNIX workstation to take advantage of its greater speed, memory, and storage capacity. Microcomputers continued to be vital to our work, however, as we used them for all data entry and editing. For the projects that followed, the PUMS of the 1850 and 1920 censuses, we maintained this arrangement, performing data entry on microcomputers and all post-entry data processing on a UNIX workstation.

When we began work on the 1880 project, the highest priority for software development was the data-entry program. Once that was in place, we developed the programs that checked for errors in the data, as well as the final recoding program. These programs took shape gradually as we assessed the quality of the data and received feedback from the research assistants who were performing the post-entry data processing. As data entry proceeded, we also constructed data dictionaries to translate alphanumeric field entries to their numeric codes. Each of these data dictionaries required software to build and maintain the file. In the final stage of the project, we developed the software that performed the numeric recoding of the alphanumeric data to produce the PUMS. We were able to carry over much of the software developed for the 1880 project to the 1850 and 1920 projects, with only minor modifications. However, in some cases—such as the 1920 data-entry program—we had to develop new software.

Data-Entry Software

The data-entry software selected for entering 1880 data was the Integrated System for Survey Analysis (ISSA). Produced by the Institute for Resource Development in association with Westinghouse, ISSA offered some advantages over other software packages. While household-level information had to be entered on a separate screen form, all the individual-level information for a dwelling could be included on the same form. This meant that our data-entry operators could see the individual-level information for the entire dwelling while entering the data for each individual. Another advantage of ISSA was that the output was in ASCII format. The census data files needed to be in ASCII format so that the post-entry data processing programs could read them.

Early on we encountered problems because we were trying to use ISSA—which had been designed to enter survey data already numerically recoded—on a project for which it had not been designed. We had determined, though, that we wanted to enter the census data in alphanumeric form as close to the original enumerator entries as possible and then perform numeric recoding as the last step in producing the PUMS. Many of the fields on the census enumeration form required that we establish large strings on the screen form to accommodate verbose enumerator entries. For example, it was not uncommon to see an entry in the occupation field such as "Works in a shoe factory." Preserving the original entry provided us with a great deal of flexibility in our method of numerically coding the data, but our record sizes were quite large and we were not able to perform many interactive data consistency checks within the data-entry software. Because ISSA was designed to accommodate pri-

marily numeric data, it offered few string-handling functions. We were able to perform only limited operations on string data within the data-entry program. We relied instead on post-entry data consistency checking software to carry out many of these functions.

We continued using ISSA for the 1850 project because we not only did not have sufficient time to carry out a study of available data-entry software but we also had already developed a workable data-entry program for the 1880 data. Because fewer questions were asked in the 1850 census than in the 1880 census, it was relatively easy to modify the 1880 version of the data-entry program to accommodate the entry of 1850 census data. Our experience in dealing with the 1880 data enabled us to avoid many problems that we encountered when we had initially set up the data-entry program. Our data-entry operators were also already familiar with ISSA, so they were able to carry over their skills to the 1850 project.

More questions were asked on the 1920 census than on the 1880 and 1850 censuses, and ISSA could not accommodate the expanded length of the 1920 individual records. We were thus forced to look elsewhere for our data-entry software for the 1920 project. We began our search for an alternative data-entry program by consulting with a variety of personnel at the University of Minnesota as well as a number of computer publications to determine what software was available. We encountered many problems similar to those we had experienced during the search for software for the 1880 project. One of the most common difficulties we faced with the database/data-entry software we tested was that many of these packages were designed to allow users to view only one record on the screen at a time while entering data. We wanted, however, to design the screen form to look as much like the enumeration form as possible, showing the maximum amount of information about the household and all individuals in the dwelling. Some Windows-based software packages, such as FileMaker Pro, are good for designing complex screen forms; but with our computers, the screen refresh rate was unacceptably slow as our data-entry operators began to enter data on the large, complicated screen forms designed to closely resemble the original census enumeration form.

Another common problem was that most database/data-entry software has its own proprietary output format, and we wanted the output to be in ASCII format. While the software packages we tested typically offered an option to convert the output to ASCII, in most cases only the file in the proprietary format could be updated. We would have had to have maintained two separate data files for every microfilm reel, one in ASCII for use in our post-entry data processing programs as well as one in the proprietary format for updating. Each of our data files is associated with a microfilm reel, and because the 1920 census consisted of more than two thousand microfilm reels, we would have had to have maintained more than four thousand files.

We decided to develop our own data-entry software using Microsoft C along with a product called C-scape by Liant Software, a library of screen-handling functions designed to allow users to create customized interactive software. Before proceeding with developing our own data-entry software, we seriously considered the disadvantages of this approach, namely the unforeseen difficulties that are part of any new software development. We nevertheless determined that we had both sufficient expertise and time necessary to accomplish the task.

Developing a customized program allowed us a great deal of flexibility in designing the screen. We designed screen forms similar to the census enumeration form, simultaneously displaying both household-level and individual-level information, and we also included customized help screens in the data-entry software. Developing our own data-entry software also allowed us to build in sophisticated interactive data consistency checking for string data as well as numeric data. By including such logic in the data-entry program, we were able to reduce the number of the post-entry changes that we had to perform.

One of the other great advantages of developing our own data-entry software was our ability to integrate the sample point selection and the data-entry procedures. In the 1880 and 1850 projects, we generated random sample point numbers in a separate program and then printed them out on a standard form. The data-entry operators marked these sheets by hand, indicating whether they accepted or rejected the sample point, along with any comments—a time-consuming process that used a great deal of paper. In the 1920 project, sample points were displayed on the screen, and the data-entry operators accepted or rejected sample points by pressing the appropriate keys. This procedure not only saved many reams of paper, but it also sped up the process and eliminated the risk of a data-entry error for the fields that uniquely identified that sample point, because those fields were automatically filled in from the sample point selection screen.

Post-Entry Data Processing

With the exception of the data-entry software, almost all the programs developed for the census projects were written for the sake of convenience in FORTRAN, a language in which we already had a number of programs, some of them obtained from the Census Bureau itself. It was quite easy to transport the programs we had developed on microcomputers to the UNIX workstation. The compilers are similar enough that only a few minor changes had to be made to the code.

Data Dictionaries

Perhaps the most important reason for moving the post-entry data processing to the UNIX workstation was to

develop large data dictionaries. A *data dictionary* is simply a computer file containing all the various entries from a field on the census form, along with the numeric recode values for each entry. We constructed data dictionaries for fields with a large number of possible entries. For some fields, such as birthplace, the data-entry operators used standard abbreviations for the most common entries, but in many cases standardization was not possible. These data dictionaries had to include all the spelling variations as well as entries infrequently appearing in the data, so the dictionaries quickly became quite large—too large, in fact, to be read into a DOS-based FORTRAN program.

We constructed data dictionaries for both household-level and individual-level fields. The fields for which we used data dictionaries in each census project are listed in table 1. An example of the contents of a data dictionary is shown in table 2, which shows a few lines for the relationship to the head-of-household field. We used data dictionaries for data consistency checking as well as for the final numeric recoding. Each line contains a field entry as it appears in the data file. In the relationship dictionary, each entry has three associated values. The first column of numbers is a general relationship code, and the next column is a detailed code. These are the numeric recode values that appeared in the final PUMS. The last column in the relationship data dictionary contains the valid entries in the sex field for the associated relationship—(M)ale, (F)emale, or (E)ither. The data consistency checking program used these entries to make sure that the entries in the sex and relationship to head-of-household fields were consistent with one another.

TABLE 1
Fields with Data Dictionaries

Household-level	Individual-level
1850	
City	Birthplace
County	Crime
State	Misfortune
Institution	Occupation
1880	
City	Birthplace
County	Occupation
State	Relationship to head of household
Institution	Sickness
1920	
City	Birthplace
County	Language
State	Occupation
Institution	Relationship to head of household

TABLE 2
Sample of the Data Dictionary for the Relationship to Head-of-Household Field

Field entry	General relationship code	Detailed relationship code	Sex
HD	01	100	E
HEAD	01	100	E
HEAD-2ND FAM	01	100	E
WIFE	02	120	F
WF	02	120	F
DAU	03	130	F
DAUGH	03	130	F
DAUGHTER	03	130	F
DTR	03	130	F
ST DAU	03	131	F
ADOPTED DAU	03	132	F
SON	03	130	M
ST SON	03	131	M
STEPSON	03	131	M
STEP SON	03	131	M
ADOPTED SON	03	132	M
ADOPT SON	03	132	M
COUSIN'S WIFE	15	243	F
ANT	15	250	F
AUNT	15	250	F

Note: M = Male; F = female; E = either.

Two programs were necessary for maintaining and updating each of the data dictionaries. The first read through all the entered census data looking for field entries that had not yet been included in the data dictionary. This program generated an output file listing all unknown entries along with the locations of those entries. A research assistant then determined the appropriate recode values for each entry and typed them into the file. A second program then read in this file and added the unknown field entries to the data dictionary along with their associated values. The program then wrote out a new version of the data dictionary sorted in whatever was deemed to be the most functional format. Some dictionaries were sorted in order of the recode values, while others were sorted in alphabetic order of the dictionary entries.

Data Consistency Checking

After the data-entry operators finished entering all the appropriate data from a microfilm reel, the file associated with that reel was put through a data consistency checking program. In the 1880 project, we carried out two data consistency checking steps separately. The first step involved checking for inconsistencies in the data, such as a male respondent having been listed as "wife" or a child of 7 having been listed as "married." This program generated a list of warning messages for a research assistant to check and determine if any of the data needed to be corrected. A sec-

ond step was later implemented to make consistency checks largely based on the contents of the relationship field. This second program performed a set of routines that linked spouses as well as parents and children. Based on these links, the software then checked for inconsistencies based on relationships—for example, if the age of a father was only ten years greater than the age of his son. We had identified a number of minor systematic changes that needed to be made to the data, and we incorporated the logic to identify these problems into the second program, so that the software could perform these changes to the data automatically.

For the 1850 data, these two programs were combined into a single program, which was not as sophisticated as the 1880 data consistency checking programs. Because the respondents in 1850 were not asked to provide their relationship to the head of the household, the spouse and parent-child linking could not be performed on the data. As with the 1880 PUMS, however, this program incorporated automatic changes to the data along with the warning messages that a research assistant then reviewed.

We decided to continue the single-program approach in the 1920 project. The output from this program indicated what changes had automatically been made to the data as well as possible inconsistencies; the program also incorporated the linking procedures we had initially employed in the 1880 data consistency checking programs.

Verification

In order to document the accuracy of the data-entry and consistency checking procedures, one out of every ten microfilm reels was selected at random and entered again by a data-entry operator different from the one who had originally entered the reel. We then read the two files into a verification program and compared the files field by field. Transferring the verification program to the UNIX workstation brought about a dramatic improvement in performance. No longer constrained by the limited memory of the microcomputers, we could read two complete data files, each of which was typically 100K or larger, into memory.

The trickiest problem in the verification program stemmed from incorrectly entered key information. The *key* is made up of the fields that uniquely identify each case, consisting of a sequence number, the page number on the enumeration form, and the line number of the first individual taken. If any of these fields were entered incorrectly, the verification program would not be able to match a case with the corresponding case in the other file. To deal with this problem, the verification program displayed a list of all keys present in one of the files but not the other. A research assistant examined these cases to determine if the key was incorrect, indicating if this were so when prompted by the program. We avoided these problems in the 1920 project because the data-entry program automatically filled in the key values when the data-entry operator accepted a sample point.

In the final step, the program created an output data file in which all incorrect responses from the original file were replaced with the correct ones. The program also generated a number of reports that allowed us to calculate our overall error rates and enabled us to identify any systematic problems. When all differences had been checked, the program produced a report indicating the error rates for each field and an itemized list of all changes made. This report also contained a list of all sample point selection errors committed by the data-entry operator. The same procedures were followed for the 1880, 1850, and 1920 census projects.

Recoding

The final step in preparing the PUMS consisted of running all the data through a final recoding program. This program converted all alphanumeric data to numeric recode values and formatted the output into household records and person records. For those fields listed in table 1, recoding was done using the data dictionaries. For other fields with a limited number of valid entries, such as marital status and sex, the recode values were "hardcoded," written into the program itself. This program read the data dictionaries into memory prior to processing the data. Because of the limited memory of the microcomputers, only one (or a part of one) dictionary could be read into memory at a time. For each data file, each dictionary had to be read into memory separately, so on the microcomputers the recoding process was enormously time consuming. Because the UNIX workstation version of this program could hold all the dictionaries in memory at one time, the program needed to read the data dictionaries only once. Whereas the microcomputers would have taken days to recode the entire PUMS, the UNIX workstation software only took about three hours to complete the entire task.

The software and hardware markets are constantly changing. At each stage of these PUMS projects, we made decisions based on the best options available to us at the time. Improvements in hardware and software often brought about dramatic improvements in productivity. The challenge we constantly faced was to make the best use of such improvements while still maintaining compatibility with what we had already developed.

PART 3. From Microfilm to Microdata: Creation of the Public Use Census Files for 1850, 1880, and 1920

Data Entry and Verification

William C. Block and Dianne L. Star

The data-entry system for the 1880, 1850, and 1920 Public Use Microdata Samples (PUMS) emphasized the dual goals of accuracy and efficiency. Achieving these objectives was not an easy task given the often poor condition of filmed manuscript census schedules as well as census takers who enumerated persons, households, and even entire districts in unorthodox ways. This essay briefly describes four issues centered around data entry and quality: the standardized system of abbreviations, symbols, and comments that went into data entry; difficulties encountered when the manuscript schedules were partially deteriorated prior to filming or when enumeration was problematic; the evolution of data entry from 1880 to 1850 to 1920; and verification and transcription error rates.

In order to move the census data efficiently from manuscript to computer, we developed a system of symbols and standardized comments for abbreviating common responses. These abbreviations ensured that data were entered using a minimum number of key strokes. Such abbreviations were used for common responses to questions regarding relationship, occupation, birthplace, marital status, etc. A second purpose of standardized abbreviations was to reduce the size of data dictionaries. As we realized partly through the 1880 project, precision in entering each enumerated spelling variation led to unwieldy data dictionaries containing hundreds of superfluous entries.

Data symbols also allowed post-entry processing to be conducted efficiently. Various symbols were designated for missing and illegible responses, illegible letters within responses, changes or additions made by the Census Bureau, and suggestions on the part of data-entry operators in cases of unclear or ambiguous data. The missing data symbol saved time during the checking process by indicating that the enumerator, rather than the data-entry operator, had neglected to enter a response. A response with illegible letter symbols might be deciphered at a later time or left intact, as often happened with names that were difficult to read. When the Census Bureau made changes or additions (e.g., more specific information about counties or cities), a symbol was entered that allowed this information to be included in the field.

Over the course of entering ten nationally representative subsamples, a set of standardized comments was developed to provide concise, consistent, and machine-readable statements for later processing. These comments indicated that certain variables contained information changed or added by the Census Bureau, responses were enumerated in the wrong columns but were still considered valid, or certain fields contained information that exceeded the length of the field in the data-entry program or contained other oddities requiring further investigation or consideration. Standard comments also made it possible to preserve extra information that went beyond what was requested in the schedule instructions (e.g., the number of months a particular child had been in school). Nonstandard comments were used to capture the oddities of census enumeration that did not fit within the framework we developed for standardized comments.

When the filmed manuscript schedules were in good condition, and when enumerators followed instructions and wrote clearly, the task of data entry was relatively straightforward. When deteriorated pages, poor handwriting, bad spelling, out-of-order pages, and idiosyncratic enumerators were encountered, data entry was made more difficult. Often, however, it was possible to record most cases with a high degree of certainty using information available on the manuscript. For example, in those instances in which enumerators failed to record dwelling and family numbers correctly, other information—such as street addresses, relationships, and even blank lines between dwellings—was usually provided and served as proxies for dwelling and family numbers. Some enumerators used their own system of ditto marks and abbreviations; some dittoed their information vertically, others horizontally. Some used dashes, and some merely entered "do." Occasionally even blank fields indicated dittoed information, although such cases

were unusual and required careful examination before data entry proceeded. Finally, from the perspective of data entry, a few variables seemed more prone to error than others. For example, some enumerators reversed first and last names; others inadvertently changed the literacy column from "cannot" read or write to "can" read and write. The unemployment variable was also prone to error, as some enumerators seemed to misunderstand the meaning of the question, listing instead the number of months employed rather than months unemployed. More systematic research has been conducted on this subject, and interested readers are referred to King and Magnuson (1993).

As data entry progressed from 1880 to 1850 and then 1920, we implemented new procedures based on our growing experience. During the first three out of ten subsamples for 1880, the data were entered exactly as they appeared on the census schedules. Such strict adherence to the original responses created unnecessarily large dictionaries and inconsistencies in the handling of odd occurrences. Beginning with the fourth subsample, the abbreviations, data symbols, and standard comments previously described were introduced. When we completed the entry of the 1880 data, project members met to discuss improvements in the data-entry phase of 1850. Several changes were made, the most important being that the missing response entry was allowed in more fields. For the 1850 project, we allowed missing field responses in the variables of city, name, birthplace, and occupation in addition to those allowed in 1880 (race, sex, and marital status). Other symbols were introduced that allowed data-entry operators to note changes made by the Census Bureau and to make suggestions in the city, occupation, and birthplace fields instead of making standard comments that were not as helpful in later stages of the project.

Before the 1920 project began, further improvements in data-entry procedure were implemented. First, the data-entry software was replaced by software designed specifically for 1920. One enhancement of the new program was automatic entry of certain variables (e.g., reel number and state). Second, more standard abbreviations were created for certain fields (e.g. birthplace and mother tongue). Third, data-entry operators no longer corrected responses when other information in the case permitted the cleaning program to impute the response, such as blank parental birthplaces when parents were present. Finally, consistency checking became a dynamic part of the program, seeking to identify inconsistent data during the data-entry process. For more information about the enhanced data-entry program for 1920, see Todd Gardner's article on software development in this issue (pp. 59–62).

As data entry progressed, we instituted a system of verification to evaluate the accuracy of the data. This verification process served three purposes. First, it provided research assistants with a means of identifying data-quality problems with specific variables and/or data-entry operators. Second, verification allowed for the creation of a transcription error rate estimate for each variable. Third, verification allowed the creation of a nationally representative subsample, containing verified information on one-tenth of the persons in the entire sample. Users who do not require the entire dataset can select this verified subsample from the final dataset.

During the 1880 data-entry project specifically, verification was conducted on two levels. The first 10 percent of the cases entered were sight verified by research assistants who verified all sample point decisions and key variables. This ensured from the beginning of data entry that sampling rules were followed correctly and data were entered consistently. Sight verification also permitted us to check closely the quality of the work during the early stages of our first data-entry project when the operators were relatively inexperienced with sampling procedures, census returns, and nineteenth-century handwriting.

TABLE 1
Transcription Error Rates from Verification, 1850 and 1880: Percentage of Error

Variable	1850	1880
CITY	1.90	0.07
COUNTY	0.83	0.09
STATE	0.00	0.21
LINE NUMBER	0.03	0.11
DWELLING NUMBER	1.49	0.62
DWELLING SIZE	0.65	0.51
NUMBER OF FAMILIES	0.74	0.15
FAMILY SEQUENCE NUMBER	0.00	0.02
FAMILY SIZE	1.10	0.59
INSTITUTION	0.68	0.15
FAMILY NUMBER	1.02	0.78
RACE	0.05	0.08
SEX	0.19	0.09
AGE	1.02	0.66
RELATION TO HEAD	NA	0.10
MARITAL STATUS	0.06	0.19
OCCUPATION	0.55	0.24
UNEMPLOYMENT	NA	0.03
SCHOOL	0.38	0.15
LITERACY	0.27	0.16
DISABILITY	0.01	NA
BIRTHPLACE	1.30	0.28
FATHER'S BIRTHPLACE	NA	0.32
MOTHER'S BIRTHPLACE	NA	0.47
SICKNESS	NA	0.06
BLIND	0.00	0.01
DEAF	0.00	0.01
IDIOTIC	0.00	0.01
INSANE	0.00	0.02
MAIMED	NA	0.02
PAUPER	0.00	NA
MISFORTUNE	0.00	NA
CRIME	0.01	NA
YEAR	0.01	NA

In addition to sight verification, 10 percent of the 1880 and 1850 microfilm reels were subjected to reentry verification. For 1880, one reel out of every ten completed reels per data-entry operator was randomly selected for reentry by a second operator. In the 1850 project, the process of selection differed slightly, although we still verified one reel in ten. Both versions of the data were compared, field by field, and a report of differences produced. A research assistant then made corrections to the original data, if required, and tallied a transcription error rate for each variable. The final error rates for 1850 and 1880 are shown in table 1. Preliminary rates indicate acceptable levels of transcription error for 1920 as well, but these are based on too few cases to publish.

The transcription error rates for both 1880 and 1850 are quite good. The total error rate for each project is less than 0.3 percent. To compare the error rates for each variable, however, it is important to understand how both substantive and procedural issues have influenced the variation in error rates between the two datasets. The variable state, for example, in 1850, was a computer-generated entry (and thus has an error rate of zero), compared with manual entry in 1880, which had an error rate of .21 percent. Variables identifying families and dwellings—dwelling number, number of families, family size, and family number—on the other hand, all contain higher error rates in 1850 than in 1880. Information on family relationships in 1880 allowed data-entry staff to resolve ambiguous cases in a more consistent fashion. The CITY variable contains a higher error rate for 1850 than 1880 because of differences in verification procedures. During 1880 data entry, when the CITY field contained an entry such as Covington, Covington Township, or Covington City, no error was assigned the original reel if the compare reel contained one of those entries. Such failure to assign errors in these situations, however, made subsequent construction of the city dictionary much more difficult and consumed many more research-assistant hours than would have been the case had the data-entry phase been more sensitive to noting differences between townships and cities, etc. This situation was corrected for the 1850 census, when such differences were carefully noted during data entry. As a result, the error rate rose dramatically in 1850 and may be viewed as a more realistic assessment of the difficulties in entering the complicated field of CITY.

Carefully devising a system of consistent and efficient data entry for the censuses of 1880, 1850, and 1920 not only required much thought but also involved a process of trial and improvement as we moved from one stage of each project to the next. Timely completion of data entry and the excellent transcription error rates indicate that we accomplished our most important objectives. Given the improvements we have implemented during each data-entry project, we are certain that the 1920 data will be of comparable or even better quality than the two preceding datasets. With the release of 1-in-100 samples of the 1880, 1850, and 1920 U.S. population, researchers will have available three excellent nineteenth- and twentieth-century individual-level datasets.

PART 3. From Microfilm to Microdata: Creation of the Public Use Census Files for 1850, 1880, and 1920

Data Consistency Checking

Daniel C. Kallgren and David Beck Ryden

Data consistency checking was integral to the construction of the 1850, 1880, and 1920 datasets. Consistency checking was a procedure designed to ferret out both data-entry errors and apparent enumeration mistakes or inconsistencies. By systematically applying a computerized checking routine to each case entered, we hoped to limit the number of errors in the final Integrated Public Use Microdata Series (IPUMS) to a minimum.

Overview

The checking process began after each reel of the microfilmed census manuscripts was coded into machine-readable form. These raw data were submitted to a consistency program that scanned each individual case for possible errors and logical inconsistencies. For example, the program was designed to detect unlikely relationship descriptions such as a 2-year-old boy who was coded as "divorced." The software also made a number of simple changes to the data automatically. The output generated by the checking software included a list of these changes and a list of problem cases that the research staff then examined individually. If the problem was the consequence of data-entry error, the research staff made the appropriate correction. If the difficulty lay with the enumerator's procedure and if the research staff was able to determine the problem, the information was then corrected. If the problem could not be resolved, the staff left the case as originally enumerated. All cases modified from the original census schedules are distinguished in the final datasets by a "data-quality flag" variable. This variable indicates which changes were made by the software and which were made manually by research staff.

The amount of work related to consistency checking depended on the quality of the microfilm and the enumeration. At times, the microfilm was difficult to read because of poor film quality. Manuscript pages were occasionally out of order, disrupting the sequence of dwelling and family numbers and, on rare occasions, dividing households. A more common problem, however, was the deterioration of the census manuscripts themselves. Some information was lost due to frayed edges and corners of the enumeration schedules. The difficulty in reading these manuscript facsimiles was at times further compounded by enumerator carelessness or ignorance of proper census-taking procedure.

The number of enumeration errors was directly correlated with the scope and complexity of the census. Of the three PUMS coded at the University of Minnesota, the 1850 manuscripts contained the fewest errors and inconsistencies since they lacked the complex "relationship to head of household" variable as well as the "marital status" variable. On the other hand, 1880 was the most problematic because the Census Bureau failed to instruct the enumerators properly on methods of recording new variables. Some census takers identified census families and dwellings incorrectly. Some enumerators reversed first and last names and sometimes reversed the unemployment variable, providing the weeks employed instead of the weeks unemployed. In most cases, we were able to determine the problem and make the appropriate corrections.

Interrelationship Checking Procedures

We used a variety of computerized checks for geographical, occupational, and data-entry inconsistencies. However, the largest number of problems came from the family relationship variable recorded in the 1880 and 1920 censuses. Unfortunately, these inconsistencies were typically too complicated to be corrected automatically by machine. We therefore designed the checking software to identify each problematic case as well as produce a message describing the nature of the inconsistency. The research staff looked up each of these cases on the microfilm and determined family relationships based on clues recorded on the manuscript. If the family relationship could not be determined, the researchers made no changes to the data.

Although there was a degree of subjectivity in solving interfamily relationship inconsistencies, the research staff

followed a set of well-defined rules before making changes to the data. Most decisions involved either inferring relationships or altering the division of dwellings into households. The basic objective in making changes to the relationship variable was to ensure that each individual was identified in relation to the household head, and the household head was listed first in any family. In achieving this goal, we encountered a variety of interrelationship problems that were resolved either by inferring family relationship, merging multiple households into one, or splitting an enumerated family into two. The following discussion details each of these modifications to the data and provides illustrative examples of particular difficulties that we encountered.

Inferring Family Relationship

Inferring family relationships was necessary when the field was left blank, or when relationships conflicted with other enumerated information. Potential conflicts were identified when children had different surnames from the head, the age difference between children and parents was too small, or the sex and family relationship was incompatible (such as a female enumerated as an uncle). Inferences were based on all other available information such as sex, age, race, marital status, birthplace, parental birthplace, and position in the family.

In figure 1, for example, George E. Morton, person 3, appears to be the son of George B. Morton because of surname, parental birthplaces, and age. After making this inference, we changed Mary A. to daughter-in-law, and the children to grandchildren so that they were properly related to the head of household.

Sometimes a child had a different surname from the head of the household because his or her mother had remarried. If this looked likely, then the children could actually be stepchildren, and we entered stepson or stepdaughter where appropriate. In about 10 percent of cases, a different surname occurred because the operator had failed to correctly indicate a reversal of the first and last names in the comment field. Even more common were slight spelling differences between the head and child. We standardized the spellings of surnames by adopting the head's spelling, unless it looked implausible. It was important to correct these two errors, because surname similarity was used for parent-child linking (see Ruggles, pp. 52–58).

Twenty percent of the time a surname difference occurred because the child was not really a child of the head but rather the child of a servant, boarder, or other relative. If the immediately preceding adult shared the surname of the child and was of a reasonable age to be a parent and there were no parental birthplace conflicts, we thought it safe to assume that he or she was the true parent and altered the child's relationship codes accordingly. For instance, in figure 2 we decided—based on surname, age, occupation, and mother's birthplace—that Annie Pipper, person 5, was the daughter of Lucie. We therefore changed the relationship from daughter to servant's daughter.

Merging Two Enumerated Households into One

When it came to identifying separate households, some enumerators were overzealous and split every dwelling into multiple families. To resolve this problem, we adopted the policy of letting the relationship field take priority over the family numbers. When the first person in the second family had a valid relationship to the head of the first family within the same dwelling, the second family number was replaced with that of the prior family. This change is indicated by the data-quality flag QHHNUM on the person record.

FIGURE 1
Relationship of Person 3 Changed to "Son" and Subsequent Relationships Changed to Match

Dwelling [6 2961 30] Rule: 1 DWSIZE: 7 NBRFAMS: 1 SEQFAM: 1 FAMSIZE: 7 NBRTAKEN: 7
(3) Probably related to HEAD
(3) Operator Comment: REL ERR; COHEAD OR SON
(4) WIFE not in second position
(4) Operator Comment: REL ERR
(5) Operator Comment: REL ERR
(6) Operator Comment: REL ERR
(7) Operator Comment: REL ERR

	FAMNO	Last Name	First Name	R	S	AGE	Relationship	M	W	Occupation	BPL	PBPL	MBPL
(1)	44	MORTON	GEORGE B	W	M	71	HEAD	M		FARMER	Mass.	Mass.	Mass.
(2)	44	MORTON	MARY A	W	F	69	WIFE	M			Maine	Maine	R.I.
(3)	44	MORTON	GEORGE E	W	M	43		M		FARMER	Mass.	Mass.	Maine
(4)	44	MORTON	MARY A	W	F	37	WIFE	M		KEEPING HOUSE	Mass.	Mass.	Vt.
(5)	44	MORTON	CLARENCE E	W	M	14	SON	S		ATTENDING SCHOOL	Ill.	Mass.	Mass.
(6)	44	MORTON	LESTER R	W	M	11	SON	S		ATTENDING SCHOOL	Ill.	Mass.	Mass.
(7)	44	MORTON	HATTIE	W	F	4	DAUGHTER	S			Ill.	Mass.	Mass.

In general, enumerators were instructed to treat servants and boarders as part of the family with whom they resided, even if they took their meals separately. However, we treated as separate units families of boarders who had their own family number. In the case of servants with their own family number, we checked the occupation of the dwelling head to see if it seemed plausible that they were servants of the head; if so, we merged them with the head's family.

Figure 3 presents a case where we merged two families. Vance Hammel was clearly the son of Crawford but is listed as the first person in the second family of the dwelling. We assumed that the family relationships are correct, which means the family numbers are wrong. To fix the problem, we changed Vance's family number to 123, changed the "number of families" variable to 1, and changed the family size to 6.

Splitting an Enumerated Family into Two

When two apparently unrelated families lived in the same dwelling, but were given the same family number, we divided the unit into two. This problem occurred when an enumerator simply forgot to change the family number within a multifamily dwelling. Before making the decision to change a single-family dwelling into a multiple dwelling, we looked closely at a number of related questions listed on the census schedule. We checked the relevant surnames, birthplaces, parental birthplaces, ages, and sometimes occupation to ensure that the secondary related group had no clear relationship to the primary family. We kept in mind the fact that many relatives such as in-laws and married daughters may have had different surnames. If we were persuaded that a secondary related group was unrelated to the head, we changed the group into a separate family by assigning a new family number.

At first glance, figure 4 looks like a classic secondary family. William and Julia Ferguson, persons 9 and 10, are just an elderly married couple residing with the widow Vincent. But notice that the Va./Ala. parental birthplaces of the head just match the Fergusons. This seems unlikely to be a coincidence. Thus we changed William to father and Julia to mother.

On the other hand, the Youmanses and the Wallaces in figure 5 indicate no relationship at all. O.V. Wallace also shows no relationship to head listed, often the case for heads of households. We therefore assigned O.V. the rela-

FIGURE 2
Relationship of Person 5 Changed to "Servant's Daughter"

Dwelling [176 241 20] Rule: 1 DWSIZE: 4 NBRFAMS: 1 SEQFAM: 1 FAMSIZE: 4 NBRTAKEN: 4
(4) Child has different surname from HEAD

	FAMNO	Last Name	First Name	R	S	AGE	Relationship	M	W	Occupation	BPL	PBPL	MBPL
(1)	92	ALLEN	ALBERT A	W	M	59	HEAD	M		FARMER	Ky.	Va.	Va.
(2)	92	ALLEN	SARAH	W	F	28	WIFE	M		KEEPING HOUSE	Ky.	Ky.	Ky.
(3)	92	ALLEN	BETTY	W	F	6	DAUGHTER	S			Ky.	Ky.	Ky.
(4)	92	PIPPER	LUCIE	B	F	35	SERVANT	S		SERVANT	Ky.	Ky.	Ky.
(5)	92	PIPPER	ANNIE	B	F	8	DAUGHTER	S			Ky.	Ky.	Ky.

FIGURE 3
Two Families Merged into One

Dwelling [76 2892 29] Rule: 1 DWSIZE: 8 NBRFAMS: 2 SEQFAM: 1 FAMSIZE: 4 NBRTAKEN: 8
(5) First position not HEAD
(7) Child has different surname from HEAD
(8) Child has different surname from HEAD

	FAMNO	Last Name	First Name	R	S	AGE	Relationship	M	W	Occupation	BPL	PBPL	MBPL
(1)	123	HAMMEL	CRAWFORD	W	M	69	HEAD	M		LABORER	Pa.	Pa.	Pa.
(2)	123	HAMMEL	PRUDENCE	W	F	61	WIFE	M		KEEPING HOUSE	Pa.	Pa.	Pa.
(3)	123	HAMMEL	NANCY	W	F	27	DAUGHTER	D			Pa.	Pa.	Pa.
(4)	123	HAMMEL	JOSEPHINE	W	F	19	DAUGHTER	S		LIVING OUT	Pa.	Pa.	Pa.
(5)	124	HAMMEL	VANCE	W	M	11	SON	S			Pa.	Pa.	Pa.
(6)	124	HAMMEL	CRAWFORD	W	M	6	SON	S			Pa.	Pa.	Pa.
(7)	124	GRUB	MARGARET	W	F	4	DAUGHTER	S			Pa.	Pa.	Pa.
(8)	124	GRUB	JOHN	W	M	1	SON	S			Pa.	Pa.	Pa.

FIGURE 4
Relationship of Persons 9 and 10 Changed to "Father" and "Mother"

Dwelling [311 5272 37] Rule: 1 DWSIZE: 10 NBRFAMS: 1 SEQFAM: 1 FAMSIZE: 10 NBRTAKEN: 10
(9) Blank REL
(10) WIFE not in second position

	FAMNO	Last Name	First Name	R	S	AGE	Relationship	M	W	Occupation	BPL	PBPL	MBPL
(1)	637	VINCENT	CATERINA O	W	F	46	HEAD	W		KEEPING HOUSE	Ala.	Va.	Ala.
(2)	637	VINCENT	BENJAMIN	W	M	21	SON	S		BOOKKEEPER	Ala.	Ala.	Ala.
(3)	637	VINCENT	LOUISA O	W	F	19	DAUGHTER	S		AT HOME	Ala.	Ala.	Ala.
(4)	637	VINCENT	JOHN K	W	M	16	SON	S		AT SCHOOL	Ala.	Ala.	Ala.
(5)	637	VINCENT	ALEXINA T	W	F	13	DAUGHTER	S		AT SCHOOL	Ala.	Ala.	Ala.
(6)	637	VINCENT	ANNA M	W	F	11	DAUGHTER	S		AT SCHOOL	Ala.	Ala.	Ala.
(7)	637	VINCENT	CHARLES E	W	M	9	SON	S		AT SCHOOL	Ala.	Ala.	Ala.
(8)	637	VINCENT	FANNIE D	W	F	7	DAUGHTER	S		AT SCHOOL	Ala.	Ala.	Ala.
(9)	637	FERGUSON	WILLIAM	W	M	78	%	M		RETIRED	Va.	Va.	Va.
(10)	637	FERGUSON	JULIA	W	F	74	WIFE	M		AT HOME	Ala.	Ala.	Ala.

FIGURE 5
Relationship of Person 5 Changed to "Head"

Dwelling [28 242 45] Rule: 1 DWSIZE: 6 NBRFAMS: 1 SEQFAM: 1 FAMSIZE: 6 NBRTAKEN: 6
(4) Blank REL
(5) WIFE not in second position
(6) Child has different surname from HEAD

	FAMNO	Last Name	First Name	R	S	AGE	Relationship	M	W	Occupation	BPL	PBPL	MBPL
(1)	315	YOUMANS	S G	W	M	58	HEAD	M		FARMER	N.Y.	N.Y.	N.Y.
(2)	315	YOUMANS	RUTH	W	F	61	WIFE	M		KEEPING HOUSE	N.Y.	N.Y.	N.Y.
(3)	315	YOUMANS	SAMUEL	W	M	17	SON	S		AT SCHOOL	Calif.	N.Y.	N.Y.
(4)	315	WALLACE	O V	W	M	44		M		CARPENTER	Ohio	Pa.	Pa.
(5)	315	WALLACE	MARY J	W	F	40	WIFE	M		KEEPING HOUSE	Scotland	Scotland	Scotland
(6)	315	WALLACE	FRANK	W	M	19	SON	S		FARMER	Colo.	Ohio	Scotland

tionship of "head" and gave a new family number for each of the Wallaces.

Conclusion

Early in the planning stages for the IPUMS, we decided not to preserve obviously erroneous information in the data. We thought the samples would be most useful to researchers if such enumerator errors were corrected. Although the changes made automatically by the computer and the changes made by the research staff followed a set of clearly defined rules, a certain amount of subjectivity was involved in all these changes. By using the data-quality flags, however, researchers can choose between data that have been altered and those that have not.

PART 3. From Microfilm to Microdata: Creation of the Public Use Census Files for 1850, 1880, and 1920

Interpreting Work: Classifying Occupations in the Public Use Microdata Samples

Matthew Sobek and Lisa Dillon

From the perspective of the data user, occupation is an invaluable variable providing information on labor-force participation, technical function, and sometimes the social relations within which a person worked. Occupation can also suggest socioeconomic position and perhaps even social class. It is the key social locator available in historical census data. From the perspective of the dataset creator, however, occupation is a colossal headache requiring more interpretation in coding than any other variable. This article describes the process involved in creating the occupation variable in the 1850, 1880, and 1920 Public Use Microdata Samples (PUMS).

In the modern census, the occupation question is addressed to everyone above a certain age. This has not always been the case. In 1850, only white and free black males at least 15 years old were asked their occupation. By the 1880 census, women were asked their occupation, and the age limit for the question had dropped to 10 years. In fact, enumerators were instructed to record the occupations of even younger children if they made a significant economic contribution to their household. In 1920—close to the historical peak of child labor—no age restriction was put on the question. In all years, the occupation question was to be filled in for every person (although "none" was the prescribed nonoccupational entry for 1920). Therefore, the data contain a great number of nonoccupational responses in addition to gainful employment. The modern definition of labor-force participation is determined by whether a person worked within a given reference week. Prior to 1940, the "gainful employment" concept was operative. This undefined concept left enumerators and respondents to interpret a person's usual work activity within the parameters suggested for particular cases in the enumerator instructions. For instance, a woman working for pay at home in 1920 was to be returned as gainfully employed if she "regularly" earned money by her work.

The Occupation Variables

The mechanics for dealing with occupation in the different PUMS created at Minnesota were fairly straightforward. The data-entry operators recorded every occupation as it was written on the original manuscript census forms (some standardized abbreviations were used). In this uncoded alphabetic form, however, occupation is nearly useless. In the case of the 1880 PUMS, the data contained over twenty thousand distinct occupation responses out of the sample of roughly five hundred thousand persons. A categorical scheme was necessary to organize and condense the thousands of different job titles encountered. The Census Bureau has classified occupations since the mid-nineteenth century, changing the scheme from decade to decade, sometimes substantially. For the 1880 and 1920 PUMS, we chose to replicate the contemporary classifications used by the Bureau in its published tabulations. The 1850 system was simply a lengthy listing rather than a classification that grouped similar occupations. Lacking a useful contemporary system, we imposed the 1880 scheme on the 1850 data. In carrying out our classification, we created a file containing each unique original alphabetic response. We then assigned numeric occupation codes and entered them beside each response. This *data dictionary* was ultimately

used to translate the manuscript answers into the codes comprising the PUMS occupation variable. We created a number of nonoccupational response categories not used by the Census Bureau in order to differentiate groups like students, retirees, and women keeping house.

Because classifications changed so much over the years, it is difficult to make historical comparisons. In general, the census occupational classification prior to 1940 did not group occupations in a manner consistent with a modern understanding of occupational structure. Aside from unhelpful groupings, many occupational categories gave the work setting (e.g., "railroad employee") rather than the specific work responsibilities or tasks (e.g., "locomotive engineer"). With the original responses as a starting point, we were not constrained by the historical occupational coding systems. Following the precedent set by earlier PUMS projects, we coded occupations from 1850 through 1920 into the 1950 occupational classification system (see Sobek's article, "The Comparability of Occupations and the Generation of Income Scores," pp. 47–51). The logic underlying the 1950 system is basically socioeconomic (professionals, clericals, craftsmen, etc.), and one with which modern social scientists are generally familiar. We entered the 1950 codes in the data dictionary alongside the 1880 codes. The codings into the historical and 1950 systems were performed independently; each was based on the original hand-enumerated manuscript response as recorded by the data-entry operators.

The 1950 classification was one level of convenience we were able to build into the 1850, 1880, and 1920 PUMS. But some historians may wish to focus on particular occupations at a finer level of detail than that offered by the Census Bureau classification. For example, researchers particularly interested in mining may wish to compare gold miners and coal miners. The Census Bureau occupational classification systems do not provide such detail, instead coding thousands of occupational titles into two or three hundred categories. A number of distinctive occupations like "prostitute" were grouped with other titles (e.g., "attendants, personal and professional service") and cannot be separated out again. However, giving researchers a complete listing of all occupations as originally recorded from the schedules would provide an unmanageable level of detail preserving meaningless distinctions such as those between "gold minors" and "ogld miners" or between "c. miners" and "col miners." To accommodate more exacting research needs while eliminating mere spelling variations, we created a supplementary detailed occupational coding scheme based on the 1950 system. The detailed occupation codes are basically addenda to the 1950 classification, extending the occupation codes from three to seven digits if read as a single field. The first three digits provide the 1950 occupation code, and the last four distinguish specific job titles while removing spelling variations.

To generate the detailed codes, we began with the completed data dictionary listing every occupational response as originally recorded in alphabetic form along with the occupation codes that we assigned. We sorted the file by our 1950 codes. A unique number was assigned to each valid variation of an occupation within a 1950 category, collapsing distinctions that were unambiguously spelling variations or abbreviations. Thus "railroad contractor" and "rr contractor" were given the same detailed code, while "railroad man" and "railroad porter" were given different codes. Many occupational responses (e.g., "railroad man" and "railroad worker") were distinguished by different detailed codes, even though they seemed logically similar. Our goal was to provide researchers with as much detail as possible in case such differences turned out to be significant. Similarly, differences in terms of grammatical structure were preserved. For example, "bookkeeper" and "bookkeeping" were given separate detailed codes. Some of these distinctions may be ephemeral or useless, but we wished to err on the side of caution.

Classification Issues

Coding occupations often proved challenging, with the degree of difficulty varying by census year. There is no surviving document detailing exactly which titles got coded into each of the 265 occupational categories in 1880.[1] For practical purposes, the titles of the 265 categories were our sole guide for 1880 coding. But even if a detailed occupational index had existed for 1880, we would still have been faced with the problems posed by the occupational responses themselves. Sometimes the response was vague, or an industry or place of work rather than a job was listed. Fortunately, most such problematic occupations were unique or contained only a handful of persons. While there were many problematic titles, the majority of persons gave responses that could be classified unambiguously.[2]

In order to ensure consistent coding, we applied a few standard rules for classification. We used the first occupation when more than one was listed. In coding into the 1880 system, we favored industry over occupation when both were listed, because the 1880 system was more oriented toward work setting than technical function. If the response referred to a shop, the person was coded within manufacturing, whereas reference to a store placed a person within trade and transportation. If the title listed only a type of store with no job description, we coded the person as a trader and dealer in that line of trade. We checked the manuscript reel for additional information in the relatively few cases in which the coding rules were insufficient to classify the occupation.

For the 1880 PUMS, we imposed the final step of comparing our results for discrepancies with the published 1880 tabulations. The comparison revealed dramatic differences for a handful of occupations. Significantly, the PUMS pro-

ject encountered 80 percent more domestic servants than the Census Office had recorded.[3] The problem lay in the great number of housekeepers in the PUMS. Our review of the responses showed that these were women unequivocally returned as "housekeepers" on the schedules. Despite specific instructions to enumerators on this point, they clearly had not accurately distinguished housekeepers from women keeping house (a nonoccupational response). On some unknown basis, census tabulators changed these women's occupations—probably at the point of initial interpretation from the manuscripts. The disparity between census and PUMS figures demonstrates the use by the Census Office of other personal, household, and location characteristics in assigning occupations, even to the point of superseding their own enumeration instructions. Given the historical trajectory of women's work, the census figures are more plausible than the unadjusted PUMS results, particularly with respect to married women's labor-force participation. The unadjusted 1880 figures would suggest an implausible, dramatic temporary spike in married women's work for pay outside of home. We imposed a logical rule recoding all housekeepers to a new nonoccupational code if the woman was related to the head of the household.

The second major adjustment we made on the basis of the published figures had to do with agricultural laborers. Over the years, the census reports contain a running commentary on the difficulty of distinguishing agricultural from other types of laborers because of the enumerators' failure to make the distinction. Consistent with the problem outlined by the Census Office, we had more laborers and fewer agricultural laborers than the published data. We recoded laborers as farm laborers if they lived in a household headed by a farmer, unless they had an explicitly nonagricultural element in their response (e.g., railroad track laborer).

The published 1880 data were also helpful in interpreting the "nurse" and "clerk" responses. With nurses, the issue was distinguishing medical nurses from domestic servants. To approximate the 1880 figures, we had to split the response "nurse" between the two occupations—something the Census Office apparently accomplished. We coded these women as domestic nurses (domestic servants) if their relationship to the household head was "resident employee."[4] Clerks did not require a similar logical recode, but the published data suggested that the simple response of "clerk" should be coded as "clerks in stores" rather than "clerks and copyists, not otherwise described." We do not know if this interpretation is, in fact, the one used by the Census Office. Where appropriate, we applied the above 1880 rules to the 1850 data as well. Since only men were asked the occupation question in 1850, only farm laborers and clerks were subject to these changes.

For all three PUMS, coding into the 1950 system was simplified greatly by the *Alphabetical Index of Occupations and Industries* (U.S. Bureau of the Census 1950), the document the Bureau used in its own tabulations of occupation from the manuscripts for the 1950 census. The index translates thousands of specific job titles into the 1950 occupation codes. The vast majority of occupational titles from the earlier PUMS years were listed in this index, which contains a great deal of historical material. We adhered strictly to the coding prescribed by the 1950 index and did not attempt our own determination of changes in the status or function of particular occupations.

The 1920 data presented fewer classification problems than did the earlier PUMS. The Census Bureau coders in Washington actually wrote the 1920 occupational codes onto the manuscript census forms, and the Minnesota data-entry operators recorded them. The codes are consistent with the published 1920 classification scheme. In addition to the handwritten codes, we obtained from the National Archives a 1920 *Occupation Index* (U.S. Bureau of the Census 1920) translating occupational titles into the 1920 categories. We had little trouble accurately replicating the Census Bureau classification for 1920. Our early work with the data has uncovered some Bureau coding errors with respect to the handwritten codes, presenting a unique opportunity for statistically analyzing historical error rates in occupation coding. We included the 1920 code as written on the manuscripts, in addition to a PUMS 1920 code differing from the original when we deemed it to be in error. Coding into the 1950 system was easier for 1920 than for the earlier PUMS. The extreme detail of the 1920 system already written on the manuscripts aided in this classification, as did the inclusion of a separate industry question in 1920. For a significant proportion of occupations, the 1920 codes translated directly into a specific 1950 category. In the rest of the cases, we resorted to the *Alphabetical Index of Occupations and Industries* (U.S. Bureau of the Census 1950) for coding purposes.

Conclusion

It is worth reinforcing the point that a fundamental difference exists between our coding procedure and that employed by the Census Office. The key difference is our reliance on the occupation field considered in isolation from any other information. Only in a few cases of severe disjuncture between our figures for 1880 and those of the Census Office did we make changes based on other characteristics of the individual or household (i.e., housekeepers, nurses, farm laborers). Our experience suggests that the Census Office regularly employed a very different coding procedure. The census tabulators coded directly from the manuscripts. As they coded, they had before them the context of the locality and family as well as the other characteristics of the individual, such as age and sex. Furthermore, comparing the published 1880 figures with our own for women, children, and the elderly (the only groups distinguished) reveals that the Census Office rejected "unlikely"

or "impossible" responses. Further work may uncover similar practices in the 1920 coding.[5]

There are some advantages to the Census Office method in terms of error control. The context offered by the manuscript returns can in some cases suggest more reasonable interpretations of the occupation response. It is also possible that most cases of women or children in unusual occupations were errors, as the Bureau concluded. The census method, however, builds correlations between characteristics into the coding scheme itself, injecting certain preconceptions and assumptions into the data. Furthermore, this practice undoubtedly changes legitimate responses for some persons with unusual occupations for their age and sex: the very kind of information that modern scholars may find most interesting. Our classification left the occupation variable independent. Even when we made changes after the fact, we provided the information necessary to undo them.

In the PUMS, we have tried to provide researchers with as much occupational information as possible. We coded each dataset into two separate systems to facilitate historical comparisons. Using the detailed occupation codes, scholars can deconstruct our categories. With respect to the 1880 coding scheme, we tried to uncover the Census Office logic and have indicated where we encountered difficulties replicating it. Our experience makes it abundantly clear, however, that the Census Office coded by the seat of its pants. We should not be naive about the sanctity of the data or the published census figures. Women's historians have rightly noted the problems associated with census counts of women's work, but there is a broader issue. There is a great deal of interpretive "play" in the data that no amount of systematization can eliminate. No magic formula for coding occupations can completely surmount the incidence of ambiguous or incomplete responses, particularly in the nineteenth century. Problematic responses are not the rule, but there are enough to leave the coding strewn with small and large-scale interpretive decisions, whose validity are not verifiable.

Despite such problems, occupation remains a tremendously useful variable that provides invaluable information. Age, sex, and race are key characteristics intrinsic to an individual. Family relationship describes a person's position within the fundamental social unit. But only occupation gives an individual's relationship to society in such a succinct and powerful way. In the 1850, 1880, and 1920 censuses, it is the key social locator.

NOTES

1. Margo Anderson provided us with a preliminary 1880 occupational listing from the National Archives. The document provided some useful hints about only a few categories, one being the expansive definition of government officials.
2. In the course of coding, we also developed some occupation-specific interpretive rules as described in the respective PUMS codebooks.
3. In 1902 the Census Office was made permanent by congressional act, and it was renamed the Census Bureau. This article uses whichever name was appropriate to the period.
4. The recoding of laborers, nurses, and housekeepers was carried out in both the contemporary and 1950 classifications. When we imposed such logical changes, we provided a flag or unique code allowing researchers to recombine the categories.
5. The practice of screening for impossible or improbable responses is documented in a number of census archival manuscript sources from the early twentieth century.

PART 3. From Microfilm to Microdata: Creation of the Public Use Census Files for 1850, 1880, and 1920

Geographic Variables in the PUMS

Ron Goeken and Matthew Mulcahy

The Public Use Microdata Samples (PUMS) for 1850, 1880, and 1920 contain two basic pieces of geographical information about individuals: place of birth and current place of residence. For all three censuses, enumerators were instructed to record the individual's birthplace by state for people born in the United States and by country for people born elsewhere. In 1880 and 1920, enumerators also recorded the birthplaces of an individual's mother and father. In coding birthplaces, we followed the system developed for the 1900 PUMS, which was based on the published materials from that year, with additional categories created as needed.

Instructions to enumerators for listing birthplace information were simple, and few failed to record the necessary information. Likewise, interpretation of the variable is straightforward in all three census years, although a few complications should be noted by researchers. In 1850, persons born in the Southwest, which later became part of the United States, often listed their birthplace as Mexico. Native American birthplaces were enumerated as a specific tribe—the Choctaw Nation, for example—or were listed as "Indian." The designation Indian Territory was used to describe the eastern half of what eventually became Oklahoma.

Again, in 1880, many southwesterners listed Mexico as their birthplace. Similarly, residents of the relatively new state of West Virginia gave their place of birth as Virginia. Somewhat more complicated was the coding scheme for German immigrants in the 1880 census. Although Germany was unified in 1871, enumerators were instructed to record the specific part of Germany where individuals were born. As a result, forty-six separate codes were needed to represent German birthplaces.

Coding of birthplaces is still in progress for the 1920 PUMS. Some problems mentioned above—the coding of birthplaces for Native Americans, German immigrants, and southwesterners of Hispanic descent—were also present in 1920. However, variables for year of immigration, immigration status (naturalized, naturalization papers, or alien), and year of naturalization will supplement place of birth for the foreign born.

The second important geographic variable was place of residence, which involves three levels of information: residence by state, county, and civil division. Coding state and county of residence was relatively straightforward. Because each microfilm reel for the 1850, 1880, and 1920 censuses contains individuals from a specific state, enumerator or data-entry errors were easily corrected for state of residence. In addition, misspellings and blank entries for county of residence were corrected by examining the county entry on previous and subsequent manuscript pages. To check for inconsistencies, we also consulted an index listing counties contained on specific microfilm reels. Both states and counties were then recoded to conform to the Inter-university Consortium for Political and Social Research (ICPSR) state and county codes.

Place of residence, however, is more complicated at the level of civil division. In contrast with other PUMS, we did not classify residents by broad categories of population. The 1900 and 1910 PUMS, for example, coded individuals into categories such as "city, 1,000–2,499." Rather, residents of incorporated municipalities in the 1850, 1880, and 1920 PUMS were assigned population totals in hundreds. In addition, all New Englanders were assigned a population total regardless of whether or not they resided in an incorporated municipality. Because townships often formed the smallest civil division in New England states and assumed the functions performed by urban governments in other parts of the country, we adopted the Census Bureau's practice of considering New England townships as incorporated municipalities. [The name Census Bureau is used through-

out this article, but readers should be aware that prior to 1902 it was called the Census Office.] Although we followed similar procedures regarding civil divisions in each PUMS, we encountered specific difficulties that researchers should note, beginning with the 1850 census.

We assigned residents of all incorporated municipalities in 1850 their respective population figures based on published census materials. These figures include the free and slave populations for southern cities, despite the fact that slaves were enumerated on a separate schedule and do not appear in the PUMS. Determining if a place was incorporated proved somewhat difficult because of the poor quality of many returns. Several states had incomplete returns for some counties, and census marshals often did not differentiate between various subdivisions within counties. As it was explained, "so imperfect is the Census of 1850 in this respect that hundreds of important towns and cities in all parts of the country, and especially in the South and West, are not even distinguished on the returns from the body of the counties in which they are situated. . ." (U.S. Census Office 1854, 192). Southern states presented specific problems related to incorporated municipalities. In the South, the Census Bureau reported that, with the exception of Arkansas, "the returns . . . do not show any complete system of subdivision of counties" (U.S Census Office 1853, 1015). Southern municipalities were therefore listed in separate tables in the published census returns, but the tables noted only those cities and towns that census officials could determine from the schedules. As a result, the published tables offered only a partial listing of incorporated places.

In order to deal with these difficulties, we also used the 1870 published materials (U.S. Census Office 1872) to determine if a place was incorporated. For example, the 1850 published returns (U.S. Census Office 1853) listed Cairo as the only subdivision in Alexander County, Illinois, without indicating if it was an incorporated municipality. The 1870 published returns, however, gave the populations of incorporated municipalities for the previous two censuses, so we were able to supplement the incomplete 1850 returns in this and similar situations.

In addition to problems related to the published returns, we also had difficulty noting the boundaries of incorporated places. It was particularly troubling when a township and incorporated village or town within the township had the same name—for example, Noblesville town in Noblesville township in Hamilton County, Indiana. In these cases, enumerators often gave the place name without indicating whether it was the town or township. We dealt with this difficulty by examining the microfilmed manuscript schedules in order to distinguish the village or town from the township. An enumerator often started a new page or gave some other indication separating the incorporated population from the rest of the return.

When Francis Walker became Superintendent of the Census Office in 1870, one of his top priorities was to improve the quality of information regarding municipalities. The published returns for 1860, as well as 1850, were, in his words, "exceedingly defective and inaccurate" (U.S. Census Office 1872, xlvi). Although the 1870 returns represented a marked improvement over the previous two censuses, the Bureau took more extensive care with the 1880 census. In that year, instructions called for enumerators to:

> Begin each township (if there be more than one in a district, borough, etc.,) with a new page. The population of villages within townships should be carefully distinguished on the schedules. The population of such villages should, in all cases, begin a new page; and when the inhabitants of a village have all been entered the remainder of the page should be left blank, except with the remark here ends the village of———(Wright and Hunt 1900).

The increased attention to detail resulted in published returns for 1880 that eliminated many problems we encountered with the 1850 returns. However, in assigning population totals to individuals at the recoding stage, we often found that enumerators had persisted in giving the place name without the designation of "village" or "township." To resolve ambiguous cases, we examined the microfilm reels whenever a municipal incorporation and nonmunicipal incorporation with the same name were located in the same county. In addition, we checked cases that had a valid entry in the street field but were initially assigned a population of zero, and cases that did not have a street entry but had a population of twenty-five hundred or greater.

The assignment of population totals to residents of incorporated municipalities is still in progress for the 1920 PUMS. However, a number of changes in the 1920 manuscript schedule alleviated some difficulties we encountered in 1850 and 1880. The most important change was the inclusion of separate blanks for incorporated place and civil division, which reduced confusion in differentiating between villages and townships sharing the same name. The schedule also provided a column for farm residence, which we used to verify residential status.

The Census Bureau introduced the farm variable in an attempt to better distinguish between the urban and rural populations or, more specifically, between rural residents who were farmers and those who were not. It was one of many attempts by the Census Bureau to achieve greater precision in its classification of urban and rural residents. Unfortunately, these attempts often created more confusion than clarity. Urban and rural are valid historical concepts regarding the U.S. population, with the rural population composed predominately of farmers and their families. Relatively low population densities and travel and communication constraints resulted in the rural population's having distinct characteristics. But with the growth of suburban areas and advancements in transportation and communication, such distinctions became harder to classify and, according to one geographer, the Census Bureau ended up defining rural as "everything that is not urban" (Hart 1994, 17).

The Census Bureau originally made the distinction between urban and rural on the basis of population of incorporated municipalities. Incorporated places—either villages, towns, boroughs, or cities that satisfied a minimum population threshold—were classified as urban places, with the remainder of the population designated as rural. Such a distinction was problematic, however, because incorporation was dependent on state laws and therefore proceeded at varying rates in different parts of the country. Nevertheless, as state-level studies of Indiana and Minnesota demonstrate, villages and towns usually incorporated when they satisfied minimum population requirements (Hart and Salisbury 1965; Hart, Salisbury, and Smith 1968).

The Census Bureau used a number of different population thresholds to distinguish between urban and rural populations. After initially using a cutoff figure of eight thousand for a historical atlas in 1874, the Bureau adopted a figure of four thousand for the 1880 census. A supplementary atlas to the 1900 census, published in 1906, reduced the urban threshold to a population of twenty-five hundred. The Census Bureau also adjusted the definition of the urban population to include all New England towns with a population of over eight thousand, four thousand, and twenty-five hundred in respective Census Bureau classification systems. It should be noted that the various urban-rural classification schemes did not represent well-planned concepts. Rather, it seems that these classifications first appeared as a convenient way to map population density in counties without large cities (the eight thousand figure) and was later modified without explanation (Truesdell 1949).

Recognizing that the central city municipal boundaries were not adequate for analyzing many population centers, the Census Bureau introduced the concept of metropolitan districts in 1910. These districts included the central city along with densely settled territory adjacent to the city. The Bureau designated districts for each census year thereafter until 1940, but the definitions changed somewhat from census to census. In order to compare the metropolitan districts over time, Warren S. Thompson (1947) attempted to provide a consistent definition of metropolitan districts for the period from 1900 to 1940. Thompson's definitions of metropolitan districts, like those of the Census Bureau, were based on municipalities. In 1950 the Census Bureau introduced the Standard Metropolitan Area (SMA), which was based on counties. In the 1920 PUMS, we will provide both the contemporary definition of the metropolitan districts as well as Thompson's more consistent metropolitan district designation. As part of the larger IPUMS project (see Ruggles, Hacker, and Sobek article, pp. 33–39), we will also provide information on metropolitan residence in all census years prior to 1950, based on a modified version of the 1950 SMA definition. Although there were only a limited number of areas in the nineteenth century that met the requirements to be considered "metropolitan," these designations offer a good opportunity for comparison of large population centers over time.

The three censuses under discussion represent a continuum regarding (1) the accuracy of published population figures; (2) the ability to accurately assign those population figures to individual residents; and (3) the meaning that those population figures reveal. Regarding the first two counts, the addition of a column for farm residence and separate fields for incorporated place and civil division makes the 1920 census the most accurate. The 1880 published returns and manuscripts presented challenges, but we are confident that population totals were assigned with a high degree of accuracy. The 1850 census proved the most difficult, in large part because of enumerator error and the poor quality of the published returns.

The meaning of the assigned population figures, however, is a debatable issue. Defining categories of *urban* and *rural*, we feel, should be the responsibility of individual researchers. It is obvious that large and medium-sized cities were undoubtedly urban in character in all years. But there were also gray areas where the population totals, even if they were accurately assigned based on published census totals, do not actually reflect the urban or rural character of residents. We can hope that researchers will use the population totals along with additional information (occupation, county-level data, and metropolitan variables) to construct their own definitions of urban/metropolitan and rural when using these data. Regardless of census year, we stress the importance of specificity when defining urban and rural status and an understanding of the issues related to population totals.

REFERENCES

American Economic Association. 1899. *The Federal Census*. New York: Macmillan.

Anderson, M. 1988. *The American census: A social history*. New Haven: Yale University Press.

Cohen, P. 1982. *A calculating people: The spread of numeracy in early America*. Chicago: University of Chicago Press.

DeBow, J. D. B. 1854. *Compendium of the seventh census*. Washington: GPO.

Duncan, O. D. 1961. A socioeconomic index for all occupations. In *Occupations and social status*, by A. Reiss et al. New York: Free Press.

Eckler, A. R. 1972. *The bureau of the census*. New York: Praeger.

Edwards, A. 1938. *A social-economic grouping of the gainful workers of the United States*. Washington: GPO.

Elman, C. 1993. Turn of the century dependence and interdependence: Roles of teens in the family economies of the aged. *Journal of Family History* 18:65–85.

Featherman, D., and R. Hauser. 1976. Prestige or socioeconomic scales in the study of occupational achievement. *Sociological Methods & Research* 4:403–22.

Folbre, N., and M. Abel. 1989. Women's work and women's households: Gender bias in the U.S. census. *Social Research* 56:545–70.

Graham, S. N. 1980. *1900 public use sample: User's handbook*. Seattle: Center for Demography and Ecology, University of Washington.

Hansen M. H., W. N. Hurwitz, and W. G. Madow. 1953. *Sample survey methods and theory*. New York: Wiley.

Hart, J. F. 1994. What is rural? Typescript, Department of Geography, University of Minnesota.

Hart, J. F., and N. E. Salisbury. 1965. Population change in middle western villages: A statistical approach. *Annals of the Association of American Geographers* 55:140–60.

Hart, J. F., N. E. Salisbury, and E. G. Smith, Jr. 1968. The dying village and some notions about urban growth. *Economic Geography* 44:343–49.

Hauser, R. 1982. Occupational status in the nineteenth and twentieth centuries. *Historical Methods* 15:111–26.

Hill, B. 1993. Women, work and the census: A problem for historians of women. *History Workshop* 35:78–94.

Hodge, R., P. Siegel, and R. Rossi. 1964. Occupational prestige in the United States, 1925–1963. *American Journal of Sociology* 70:286–302.

Holt, W. S. 1929. *The bureau of the census: Its history, activities and organization*. Washington: Brookings.

Horan, P. 1978. Is status attainment research atheoretical? *American Sociological Review* 43:534–41.

Jenkins, R. 1985. *Procedural history of the 1940 census of population and housing*. Madison: University of Wisconsin Press.

Kallgren, D. Forthcoming. *Schooling in the United States, 1850–1950: The emergence of a national experience*. Ph. D. diss., University of Minnesota.

King, M. L., and S. Preston. 1990. Who lives with whom? Individual versus household measures. *Journal of Family History* 15:117–32.

King, M., and D. Magnuson. 1993. A procedural history of the enumeration process in the 1880 U.S. census. Appendix B, *1880 Public use sample*. Minneapolis: Social History Research Laboratory, University of Minnesota.

———. 1995. Perspectives on historical U. S. census undercounts. *Social Science History*.

Kish, L. 1965. *Survey sampling*. New York: Wiley.

Landale, N., and S. Tolnay. 1991. Group differences in economic opportunity and the timing of marriage: Blacks and whites in the rural South, 1910. *American Sociological Review* 56: 33–45.

Laslett, P. 1972. Introduction. In *Household and family in past time*, edited by P. Laslett and R. Wall. Cambridge: Cambridge University Press.

Magnuson, D. Forthcoming. *The making of a modern census: U.S. population census, 1850–1940*. Ph. D. diss., University of Minnesota.

McKee, O. Jr. 1930. Counting heads in the nation. *North American Review* 229:485.

Merriam, W. 1902. *Population*. Washington: GPO.

Moen, J. 1988. From gainful employment to labor force: Definitions and a new estimate of work rates of American males, 1860 to 1980. *Historical Methods* 21:149–59.

Nam, C., and M. Powers. 1983. *The socioeconomic approach to status measurement*. Houston: Cap and Gown Press.

Penn, R. 1975. Occupational prestige hierarchies: A great empirical invariant? *Social Forces* 54:352–64.

Porter, R. 1892. *Compendium of the eleventh census, 1890*. Washington: GPO.

Preliminary Inventories. 1964. *No. 161, Records of the Bureau of the Census*, compiled by K. H. Davidson and C. M. Ashby, 68. Washington: National Archives.

Riche, M. R. 1993. The Riche report. *Population Today* 21.

Ruggles, S. 1987. *Prolonged connections: The rise of the extended family in nineteenth century England and America*. Madison: University of Wisconsin Press.

———. 1991a. Comparability of the public use files of the U.S. census of population. *Social Science History* 15: 123–58.

———. 1991b. Integration of the public use files of the U.S. census of population, 1880–1980. *American Statistical Association Proceedings of the Social Statistics Section*, 365–70. Washington: American Statistical Association.

———. 1993. Historical demography from the census: Applications of the American census microdata files. In *Old and new methods in historical demography*, edited by D. Reher and R. Schofield, 383–93. Oxford: Oxford University Press.

———. 1994a. The transformation of American family structure. *American Historical Review* 99:103–28.

———. 1994b. The origins of African-American family structure. *American Sociological Review* 59:136–51.

———. Forthcoming. Living arrangements of the elderly in America, 1880–1980. In *Aging and generational relations: Historical and cross-cultural comparisons*, edited by T. K. Hareven.

Ruggles, S., and R. R. Menard. 1990. A public use sample of the 1880 census of population. *Historical Methods* 23:104–15.

Ruggles, S., et al. 1993. *1850 Public Use Microdata Sample: Preliminary user's guide*. Minneapolis: Social History Research Laboratory, University of Minnesota.

———. 1994. *1880 Public Use Microdata Sample: User's guide and technical documentation*. Minneapolis: Social History Research Laboratory, University of Minnesota.

Ryden, D. 1994. *A century of home ownership*. Paper presented at the Social Science History Association Conference, Atlanta, October.

Seaton, C. 1883. Compendium of the tenth census, 1880. Washington: GPO.

Sharpless, J. B., and R. M. Shortridge. 1975. Biased underenumeration in census manuscripts: Methodological implications. *Journal of Urban History* 1:409–39.

Smith, D. S. 1992. The meanings of family and household: Change and continuity in the mirror of the American census. *Population and Development Review* 18:421–56.

Smuts, R. 1960. The female labor force: A case study in the interpretation of historical statistics. *Journal of the American Statistical Association* 55:71–79.

Sobek, M. 1991. Occupation, class and socioeconomic status in historical census data. Paper presented at the Social Science History Association Conference, New Orleans.

Strong, M. A., S. H. Preston, A. R. Miller, M. Hereward, H. R. Lentzner, J. R. Seaman, H. C. Williams. 1989. *User's guide public use sample. 1910 United States census of population*. Philadelphia: Population Studies Center, University of Pennsylvania.

Thernstrom, S. 1963. *Poverty and progress: Social mobility in a nineteenth-century city*. Cambridge: Harvard University Press.

———. 1973. *The other Bostonians: Poverty and progress in the American metropolis, 1880–1970*. Cambridge: Harvard University Press.

Thompson, W. S. 1947. Population, the growth of metropolitan districts in

the United States, 1900–1940. Washington: GPO.
Treiman, D. 1976. A standard occupational prestige scale for use with historical data. *Journal of Interdisciplinary History* 7:283–304.
Truesdell, L. 1949. *The development of the rural-urban classification in the United States*. U.S. Bureau of the Census. Current Population Reports, ser. P-23, no. 1. Washington: GPO.
2000 census outlook: A "postcard census" it's not *Population Today* 21(12): 10.
U.S. Bureau of the Census. 1910. *Report of the director to the secretary of commerce and labor: Concerning the operations of the bureau for the year 1908–09*. Washington: GPO.
———. 1911. *Report of the director to the secretary of commerce and labor: Concerning the operations of the bureau for the year 1909–10*. Washington: GPO.
———. 1919a. *Department of commerce, annual report of the director of the census to the secretary of commerce, 30 June 1919*. Washington: GPO.
———. 1919b. *Instructions to supervisors of census: Test of applicants for appointment as enumerators*. Washington: GPO.
———. 1920? *Supplemental occupations designations for the fourteenth census—occupation index*. National Archives; Record Group 29; Box: Population Division Classification of Occupations, 1910–1917; File: 1917 Alphabetic Index of Occupations.
———. 1921. *14th census of the United States taken in the year 1920, Volume I*, Washington: GPO.
———. 1929. *Department of commerce, bureau of the census annual report of the director of the census to the secretary of commerce, 30 June 1929*. Washington: GPO.
———. 1940. *Instructions to district supervisors, population, agriculture, irrigation and administrative procedure*. Washington: GPO.
———. 1950. *Alphabetical index of occupations and industries*. Washington: GPO.
———. 1955. *The 1950 censuses—how they were taken*. Washington: GPO.
———. 1956. *Special reports. Occupational characteristics*. Washington: GPO.
———. 1961. *U.S. censuses of population and housing, 1960: Principal data-collection forms and procedures*. Washington: GPO.
———. 1966. *The 1960 censuses of population and housing: Procedural history*. Washington: GPO.
———. 1968. *Changes between the 1950 and 1960 occupation and industry classifications*. Technical paper 18, prepared by J. A. Priebe. Washington: GPO.
———. 1972a. *Public use samples of basic records from the 1970 census: Description and technical documentation*. Washington: GPO.
———. 1972b. *1970 Occupation and industry classifications in terms of their 1960 occupation and industry elements*, by J. A. Priebe, J. Heinkel, and S. Greene. Technical paper 26. Washington: GPO.
———. 1973. *Technical documentation for the 1960 public use sample*. Washington: GPO.
———. 1976. *U.S. census of population and housing: 1970 procedural history*. Washington: GPO.
———. 1983. *Public use samples of basic records from the 1980 census: Description and technical documentation*. Washington: GPO.
———. 1984a. *Census of population, 1940: Public use sample technical documentation*. Washington: GPO.
———. 1984b. *Census of population, 1950: Public use sample technical documentation*. Washington: GPO.
———. 1986–1989. *Census of population and housing (1980): History, 1980 census of population and housing*. Washington: GPO.
———. 1989a. *The relationship between the 1970 and 1980 industry and occupation classification systems*. Technical paper 59, prepared by P. Vines and J. A. Priebe. Washington: GPO.
———. 1989b. *200 years of U.S. census taking: Population questions and housing questions, 1790–1990*. Washington: GPO.
———. 1993. *Census of population and housing, 1990. Public use microdata samples: Technical documentation*. Washington: GPO.
U.S. Census Office. 1853. *The seventh census of the United States*. Vol. 1. Washington: R. Armstrong, Public Printer.
———. 1854. *Compendium to the seventh census of the United States*. Vol. 4. Washington: R. Armstrong, Public Printer.
———. 1872. *The ninth census of the United States*. Vol. 1. Washington: GPO.
———. 1882. *Report of the Superintendent of the Census (November 1, 1881)*. Washington: GPO.
U. S. Department of Commerce. 1947. *Population: The growth of metropolitan districts in the United States, 1900–1940*. Washington: GPO.
U. S. Department of Commerce. See also publications under U. S. Bureau of the Census.
Walker, F. 1883. *Compendium of the tenth census*, 1880. Washington: GPO.
———. 1888. The eleventh census of the United States. *Quarterly Journal of Economics* 2:136–61.
Wright, C., and W. Hunt. 1900. *The history and growth of the United States census*. Washington: GPO.