ACCURACY OF CENSUS DATA AS MEASURED BY THE 1970 CPS-CENSUS-IRS MATCHING STUDY

Paula Schneider and Joseph Knott
Social and Economic Statistics Administration, Bureau of the Census

INTRODUCTION

Among the various projects designed to evaluate the quality of statistics from the 1970 Census of Population is the CPS-Census-IRS Matching Study. The data in this study are based on case-by-case comparisons of persons enumerated in the March 1970 Current Population Survey and in the census 20-percent sample with regard to classification by a variety of demographic, social, and economic characteristics. In addition, for a portion of the universe the census income data were compared with information as recorded on 1969 Federal income tax returns. This paper deals briefly with the methods and design of the study and presents a capsulized view of the quality of various characteristics, as measured by the index of inconsistency and the net difference rate. All figures and statements contained herein are, however, preliminary and subject to revision. Final detailed statistics on these subjects will be published in the report series covering the Evaluation and Research Program of the 1970 Census of Population and Housing [PHC(E)].

The universe for this study was restricted to persons enumerated as members of households in the March 1970 CPS and in decennial census 20-percent households. To accomplish this match of identical persons from the two sources, the following operations were undertaken. A 1970 census geographic identification (i.e. enumeration district) was assigned to each housing unit in the March 1970 CPS and a search was made for the unit in the appropriate census address register. If an address match was made to a census unit listed as having been enumerated on a 100-percent or "short form" questionnaire, the CPS unit was dropped from the scope of this study although it was included for an associated study designed to measure population coverage in the 1970 census. If the CPS address was matched to a census 20-percent sample household, the census questionnaire was obtained and a name match performed. The matched cases constituted the universe for this study.

For each matched unit, selected CPS and census identification information on the household and persons therein was transcribed to a specially designed, machine-readable control sheet and then transferred to computer tape. This control tape was matched against the March 1970 CPS data file and to the census sample data file from which records for the appropriate persons were obtained.

Since the information was obtained from the final edited CPS and census data files, comparisons between the data sources can be used to estimate errors in publication level statistics. However, since the data are, by definition, restricted to persons for whom records were located in both sources, the measures of error reflect only content differences and not any error due to coverage problems.

In determining the levels of response error in census statistics by comparing the census classification with the corresponding classification in the Current Population Survey (CPS), the CPS response is assumed to be more accurate for some characteristics. This assumption is made since the CPS is a monthly national survey which utilizes a staff of full-time, experienced interviewers and is conducted under more extensive controls and training procedures. However, there are certain limitations involved in estimating response error by this method. First, in such a study it is seldom possible to locate the data records for all persons in both sources. In the 1970 CPS-Census Match, we were able to obtain matched records for about 75 percent of persons in the sample. If the response error distributions of the unmatched cases were generally different from those for the matched population, the data would be biased. For the purpose of the analysis in this report the assumption was made that they did not differ to an appreciable degree. Second, even though the CPS response is usually assumed to be the standard of accuracy, the CPS is obviously subject to some degree of error. In fact, for some characteristics, such as age, the CPS may be as error prone as the census. Third, whereas the CPS data are obtained through personal interview, the census data are based partially on self-enumeration responses and partially on personal interview. Therefore, differences in the type of enumeration and in the household member(s) being interviewed or completing the questionnaire may have had some effect on the responses given. The final factor to be considered in interpreting differences between the CPS and census data is the variation in time of enumeration. The census period generally extended from the last week in March and several weeks during April 1970, or longer in some areas, as compared with the 1-week enumeration period (March 16 to 20) for CPS. Therefore, some of the differences observed are correct in the sense that changes in status occur over time.

MEASURES OF RESPONSE ERROR

Two measures of response error have been estimated for each subject characteristic. These measures are the index of inconsistency - a measure of gross error or response variability - and the net difference rate - a measure of net error or bias. Estimated values of the index of inconsistency have been computed for each category in a distribution and a weighted average of the individual indices, the L-fold index of inconsistency, provides an estimate of the overall consistency of classification between the CPS and census for a given characteristic. Also, in table 1 the number of categories in the distribution from which the L-fold index was derived is shown in parentheses following each characteristic. For the purpose of measuring the adequacy of the census data collection system for a particular characteristic, an index under 20 can be considered to indicate a relatively low level of inconsistency, those between 20 and 50 a moderate level, and those over 50 a high level of inconsistency. In terms of published cross-tabulations of census data, the index

Table 1.—L-FOLD INDEX OF INCONSISTENCY FOR SELECTED POPULATION CHARACTERISTICS FROM THE C-S-CENSUS MATCHING STUDY: 1970 AND 1960

Characteristic (number of categories in distribution)	1970 L-fold index of inconsistency 1/	1970 95% confidence interval on index of inconsistency	1960 L-fold index of inconsistency 1/
AGE (17)			
Total	7	6.8 to 7.5	6
Male	7	6.1 to 7.1	5
Female	8	7.0 to 8.1	6
White	7	6.2 to 7.0	5
Negro 2/	12	10.9 to 14.1	11
Metropolitan-central city	9	7.9 to 9.5	(NA)
Metropolitan-outside central city	6	5.6 to 6.8	(NA)
Nonmetropolitan-urban	6	5.2 to 7.1	(NA)
Nonmetropolitan-rural	7	6.6 to 8.4	(NA)
SEX 3/ (2)	2	1.8 to 2.4	3
RACE (3)	3	2.8 to 4.0	4
HOUSEHOLD RELATIONSHIP			
Male (5)	4	3.6 to 4.6	4
Female (6)	5	4.1 to 5.1	5
MARITAL STATUS (5)			
Male, 14 years and over	5	4.0 to 5.5	6
Female, 14 years and over	5	4.3 to 5.6	5
EDUCATIONAL ATTAINMENT (13)			
Male, 14 years and over	38	37.2 to 39.8	(NA)
Female, 14 years and over	37	35.9 to 38.3	(NA)
VETERAN STATUS AND PERIOD OF SERVICE 4/ (6)	15	13.6 to 16.3	(NA)
EMPLOYMENT STATUS 5/ (4)			
Male	19	17.9 to 20.9	19
Female	20	19.1 to 21.7	20
WORK EXPERIENCE 6/ (7)			
Male	43	41.6 to 45.2	43
Female	37	35.9 to 38.8	36

Characteristic (number of categories in distribution)	1970 L-fold index of inconsistency 1/	1970 95% confidence interval on index of inconsistency	1960 L-fold index of inconsistency 1/
CLASS OF WORKER 7/ (7)			
Male	16	14.3 to 17.7	(NA)
Female	18	15.7 to 20.6	(NA)
MAJOR OCCUPATION 7/ (12)			
Male	31	29.4 to 32.3	25
Female	21	19.3 to 22.8	15
MAJOR INDUSTRY 7/ (12)			
Male	17	16.2 to 18.6	15
Female	14	12.3 to 15.2	10
PERSONS TOTAL MONEY INCOME (15)			
Total	45	44.0 to 45.7	38
Male	50	48.6 to 51.1	46
Female	43	41.9 to 44.5	33
WAGE OR SALARY INCOME (14)			
Male	39	38.0 to 40.5	(NA)
Female	33	31.5 to 34.2	(NA)
NONFARM SELF-EMPLOYMENT INCOME (15)			
Male	57	52.8 to 61.9	(NA)
Female	66	57.7 to 75.0	(NA)
FARM SELF-EMPLOYMENT INCOME (15)			
Male	48	42.5 to 53.4	(NA)
Female	91	68.4 to 100.0	(NA)
SOCIAL SECURITY INCOME (7)			
Male	24	21.9 to 26.8	(NA)
Female	30	27.5 to 32.3	(NA)
PUBLIC ASSISTANCE INCOME (7)			
Male	52	44.2 to 61.7	(NA)
Female	45	39.7 to 51.3	(NA)
ALL OTHER INCOME (15)			
Male	59	57.2 to 61.4	(NA)
Female	57	54.4 to 59.8	(NA)

NA Not available.

1/ The level of the L-fold index is sensitive to the detail of the classification system. For example, if age data were classified in one-year intervals, we would expect to observe more differences between trials and to obtain a higher estimated L-fold index than if the data were classified in five-year intervals. The indices shown here would not apply to published distributions where the data were either shown in more detailed or in broader categories.

2/ Refers to "Negro and other races" for 1960.

3/ Since there are only two categories in the distribution, the index of inconsistency is not an average measure.

4/ Based on civilian males 16 years old and over.

5/ Based on civilian population 14 years old and over.

6/ Based on civilian population 16 years old and over.

7/ Based on persons 16 years old and over, employed in CPS and census.

Table 2.—INDEX OF INCONSISTENCY AND NET DIFFERENCE RATE FOR SELECTED POPULATION CHARACTERISTICS FROM THE 1970 CPS-CENSUS MATCHING STUDY

Characteristic	Index of inconsistency	95 percent confidence interval on index of inconsistency	Percent in class (CPS)	Net difference rate[1/]	95 percent confidence interval on net difference rate
AGE					
Under 1 year	6	4.7 to 7.3	3.2	0.0*	-0.1 to 0.0
1 to 4 years	7	6.2 to 8.5	5.3	-0.1*	-0.2 to 0.1
5 to 9 years	5	3.9 to 5.3	10.5	0.0*	-0.1 to 0.2
10 to 14 years	4	3.5 to 4.7	11.1	-0.1*	-0.2 to 0.0
15 to 19 years	5	4.0 to 5.4	8.9	0.0*	-0.1 to 0.1
20 to 24 years	6	5.4 to 7.4	6.2	0.1*	-0.1 to 0.2
25 to 29 years	7	5.6 to 7.6	6.5	0.0*	-0.1 to 0.1
30 to 34 years	9	8.2 to 10.7	5.7	0.0*	-0.1 to 0.2
35 to 39 years	9	7.7 to 10.2	5.9	-0.1*	-0.2 to 0.1
40 to 44 years	9	7.6 to 9.9	6.1	0.0*	-0.2 to 0.1
45 to 49 years	8	7.0 to 9.3	6.2	0.0*	-0.1 to 0.2
50 to 54 years	8	7.3 to 9.7	5.8	0.0*	-0.1 to 0.1
55 to 59 years	9	7.6 to 10.2	5.1	0.0*	-0.1 to 0.1
60 to 64 years	10	8.9 to 12.0	4.1	0.0*	-0.1 to 0.1
65 to 69 years	12	9.9 to 13.5	3.5	-0.1*	-0.2 to 0.1
70 to 74 years	10	8.5 to 12.3	2.8	0.1*	0.0 to 0.2
75 years and over	8	6.5 to 9.6	3.3	0.1*	0.0 to 0.2
SEX					
Male	2	1.8 to 2.4	48.0	0.0*	-0.2 to 0.1
Female	2	1.8 to 2.4	52.0	0.0*	-0.1 to 0.2
RACE					
White	3	2.8 to 4.0	90.5	-0.3	-0.4 to -0.2
Negro	1	1.0 to 1.8	8.7	0.1*	0.0 to 0.1
Other races	19	15.5 to 24.2	0.9	0.2	0.1 to 0.3
HOUSEHOLD RELATIONSHIP					
MALE					
(Primary) family head	2	1.7 to 2.5	49.5	0.0*	-0.2 to 0.2
Primary individual	11	9.1 to 14.4	3.3	0.0*	-0.2 to 0.2
Child	2	1.9 to 2.8	43.5	0.0*	-0.3 to 0.2
Other relative	20	16.5 to 24.1	2.8	0.0*	-0.2 to 0.2
Nonrelative	29	22.4 to 37.7	1.0	0.1*	-0.1 to 0.2
FEMALE					
(Primary) family head	13	11.3 to 15.8	5.2	-0.1*	-0.4 to 0.1
Primary individual	4	3.4 to 5.5	7.5	0.0*	-0.1 to 0.1
Wife	2	1.8 to 2.6	44.6	-0.2*	-0.4 to 0.0
Child	2	1.8 to 2.1	37.8	-0.2*	-0.4 to 0.0
Other relative	18	15.4 to 21.2	3.9	0.3*	0.0 to 0.5
Nonrelative	27	21.3 to 34.9	1.0	0.3*	0.1 to 0.4
EMPLOYMENT STATUS—Con.					
FEMALE, 14 YEARS AND OVER					
Employed in agriculture	63	48.1 to 82.8	0.6	-0.2*	-0.4 to 0.0
Employed in nonagricultural industries	17	15.6 to 18.3	39.2	-1.1	-1.7 to -0.5
Unemployed	65	56.7 to 74.4	2.1	0.0*	-0.3 to 0.4
Not in labor force	19	17.8 to 20.5	58.2	1.2	0.6 to 1.9
WORK EXPERIENCE IN 1969					
MALE, 16 YEARS AND OVER					
50 to 52 weeks	32	30.2 to 33.9	64.3	-6.0	-7.0 to -5.0
48 to 49 weeks	82	74.6 to 90.8	2.6	2.6	1.9 to 3.2
40 to 47 weeks	72	66.4 to 79.2	4.5	2.2	1.5 to 2.8
27 to 39 weeks	69	62.1 to 75.7	4.3	0.7	0.1 to 1.3
14 to 26 weeks	61	55.2 to 68.2	4.8	-0.6	-1.2 to -0.1
13 weeks or less	53	47.4 to 58.6	5.3	0.2*	-0.4 to 0.8
Did not work in 1969	22	20.0 to 24.6	14.2	1.0	0.5 to 1.6
FEMALE, 16 YEARS AND OVER					
50 to 52 weeks	31	28.7 to 32.6	27.7	-5.5	-6.3 to -4.7
48 to 49 weeks	79	69.7 to 88.5	1.6	1.4	1.0 to 1.9
40 to 47 weeks	74	67.3 to 80.7	3.5	1.7	1.2 to 2.3
27 to 39 weeks	65	59.7 to 71.2	5.2	0.2*	-0.4 to 0.8
14 to 26 weeks	57	52.1 to 62.2	6.2	-0.2*	-0.8 to 0.4
13 weeks or less	50	46.2 to 54.5	8.1	0.1*	-0.5 to 0.7
Did not work in 1969	19	17.3 to 20.1	47.7	2.2	1.5 to 2.9
CLASS OF WORKER					
Agriculture:					
Wage and salary	33	26.0 to 42.2	1.2	0.4	0.2 to 0.6
Self-employed	17	12.9 to 21.9	2.3	-0.1*	-0.3 to 0.1
Unpaid family worker	53	27.5 to 100.0	0.1	-0.1*	-0.2 to 0.0
Nonagricultural industries:					
Private wage and salary	16	14.3 to 17.2	72.9	0.5*	-0.1 to 1.1
Government	12	10.9 to 13.9	16.8	-0.2*	-0.6 to 0.2
Self-employed	23	20.3 to 26.8	6.2	-0.3*	-0.7 to 0.1
Unpaid family worker	47	32.7 to 67.7	0.5	-0.1*	-0.3 to 0.0
OCCUPATION					
Professional, technical, and kindred workers	20	18.1 to 22.1	15.0	0.7	0.2 to 1.2
Managers and administrators, except farm	39	35.5 to 42.1	11.6	-2.8	-3.4 to -2.2
Sales workers	30	26.9 to 34.1	5.9	1.1	0.7 to 1.6

MARITAL STATUS					
MALE, 14 YEARS AND OVER					
Married, except separated	2	1.9 to 3.0	70.4	0.0*	-0.3 to 0.2
Separated	43	33.8 to 56.0	1.0	0.1*	-0.2 to 0.3
Widowed	14	10.9 to 19.2	2.5	-0.1*	-0.3 to 0.1
Divorced	24	18.8 to 31.1	1.7	0.4	0.1 to 0.6
Single	2	1.6 to 2.1	24.5	-0.3	-0.5 to -0.1
FEMALE, 14 YEARS AND OVER					
Married, except separated	2	1.7 to 2.6	62.1	0.0*	-0.2 to 0.3
Separated	29	23.7 to 36.3	1.8	0.1*	-0.2 to 0.3
Widowed	6	5.0 to 7.4	11.8	-0.2*	-0.4 to 0.1
Divorced	19	16.0 to 23.3	3.5	0.2*	0.0 to 0.5
Single	2	1.9 to 3.1	20.8	-0.1*	-0.3 to 0.1
EDUCATIONAL ATTAINMENT					
Elementary: 0 to 4 years	34	30.8 to 37.8	3.6	0.2*	0.0 to 0.5
5 years	67	59.3 to 75.3	1.4	0.1*	-0.1 to 0.3
6 and 7 years	46	43.1 to 49.0	6.9	0.6	0.2 to 1.0
8 years	38	36.1 to 40.3	13.3	-0.4*	-0.8 to 0.1
High school: 1 year	47	43.8 to 49.8	7.1	0.4*	0.0 to 0.8
2 years	46	42.9 to 48.4	8.4	0.6	0.1 to 1.0
3 years	48	44.8 to 51.0	6.4	1.1	0.7 to 1.5
4 years	26	25.0 to 27.3	32.8	-2.2	-2.8 to -1.7
College: 1 year	50	45.8 to 53.7	4.2	0.3*	-0.1 to 0.6
2 years	41	37.7 to 44.7	4.8	-0.4	-0.7 to -0.1
3 years	50	44.1 to 55.9	1.7	0.3	0.1 to 0.5
4 years	27	24.2 to 29.4	5.8	-0.8	-1.0 to -0.5
5 years or more	21	18.7 to 24.3	3.5	0.3	0.1 to 0.5
VETERAN STATUS					
Vietnam conflict	22	18.9 to 26.8	5.7	0.8	0.4 to 1.3
Korean conflict	20	16.7 to 23.0	8.2	1.0	0.5 to 1.5
World War II	11	9.6 to 12.9	20.2	0.2*	-0.3 to 0.8
World War I	18	13.4 to 25.0	2.2	0.3*	0.0 to 0.5
Other service	32	27.5 to 36.5	7.1	-0.8	-1.3 to -0.2
Nonveteran	10	8.3 to 10.8	56.6	-1.6	-2.2 to -0.9
EMPLOYMENT STATUS					
MALE, 14 YEARS AND OVER					
Employed in agriculture	26	22.1 to 30.4	4.7	-0.5	-0.9 to -0.2
Employed in nonagricultural industries	15	13.6 to 16.4	68.7	-0.9	-1.5 to -0.3
Unemployed	58	51.1 to 66.1	3.1	-0.2*	-0.6 to 0.3
Not in labor force	17	15.4 to 18.6	23.5	1.5	1.0 to 2.1
Clerical and kindred workers	23	20.7 to 24.6	18.1	0.2*	-0.4 to 0.8
Craftsmen and kindred workers	30	27.4 to 32.6	13.4	0.4*	-0.2 to 1.0
Operatives, except transport.	25	22.3 to 27.0	12.7	0.8	0.3 to 1.3
Transport equipment operatives	26	21.8 to 30.3	3.6	0.4	0.1 to 0.7
Laborers, except farm	49	44.0 to 54.9	4.7	-0.5*	-0.9 to 0.0
Farmers and farm managers	13	9.7 to 17.8	2.3	-0.2	-0.4 to -0.1
Farm laborers and foremen	27	20.2 to 36.7	1.0	0.1*	-0.1 to 0.3
Service workers, except private households	19	17.0 to 21.6	10.1	0.0*	-0.4 to 0.5
Private household workers	15	10.7 to 20.6	1.7	-0.2*	-0.4 to 0.0
INDUSTRY					
Agriculture, forestry, and fisheries	14	11.1 to 17.4	3.7	0.1*	-0.1 to 0.4
Mining	27	18.9 to 38.2	0.7	0.1*	0.0 to 0.3
Construction	21	18.3 to 24.6	5.9	-0.1*	-0.4 to 0.3
Manufacturing	14	13.2 to 15.9	27.8	-1.4	-1.9 to -0.8
Transportation, communications, and other public utilities	14	11.4 to 16.0	7.1	-0.1*	-0.4 to 0.3
Wholesale and retail trade	19	17.4 to 21.0	18.2	1.2	0.7 to 1.8
Finance, insurance, and real estate	11	8.8 to 13.5	5.4	-0.1*	-0.3 to 0.2
Business and repair services	34	29.3 to 40.6	2.7	0.5	0.1 to 0.8
Personal services	15	11.9 to 17.8	4.7	-0.3	-0.6 to -0.1
Entertainment and recreation services	38	27.2 to 51.8	0.6	0.1*	-0.1 to 0.2
Professional and related services	10	8.8 to 11.5	17.3	-0.1*	-0.5 to 0.2
Public administration	15	12.3 to 17.5	5.9	0.0*	-0.3 to 0.3
PERSONS INCOME IN 1969					
Without income	25	24.1 to 26.9	22.1	1.4	0.9 to 1.9
Loss	78	55.6 to 100.0	0.2	-0.1	-0.2 to -0.1
$1 to $999	47	45.0 to 49.3	14.2	-1.1	-1.6 to -0.6
$1,000 to $1,999	51	48.0 to 53.5	9.7	-0.2*	-0.7 to 0.3
$2,000 to $2,999	55	51.6 to 58.2	7.0	-0.2*	-0.6 to 0.2
$3,000 to $3,999	55	51.5 to 58.4	6.3	0.1*	-0.3 to 0.5
$4,000 to $4,999	53	49.2 to 56.4	5.7	-0.1*	-0.5 to 0.3
$5,000 to $5,999	55	51.1 to 58.5	5.4	0.2*	-0.2 to 0.6
$6,000 to $6,999	55	51.4 to 59.1	5.1	-0.1*	-0.4 to 0.3
$7,000 to $7,999	54	50.0 to 57.9	4.8	-0.2*	-0.6 to 0.1
$8,000 to $8,999	53	49.2 to 57.7	4.0	0.0*	-0.3 to 0.4
$9,000 to $9,999	53	48.1 to 57.6	3.1	0.1*	-0.2 to 0.4
$10,000 to $14,999	35	32.7 to 37.6	8.3	-0.2*	-0.6 to 0.2
$15,000 to $24,999	37	33.1 to 41.0	3.1	0.2*	-0.1 to 0.4
$25,000 or more	40	33.8 to 47.8	0.9	0.3	0.1 to 0.4

* Indicates that the net difference rate is not significantly different from zero at the 95 percent confidence level.

1/ A positive value means there were more persons in the census than in the CPS in a given category. A negative value means there were fewer persons in the census than in the CPS in a given category.

provides an approximate measure of the distorting effect each variable has upon the cross-classification. If, then, any of the characteristics have a high index value, the cross-classification may be seriously distorted. The net difference rate estimates the level of bias in the particular census distribution, where the CPS classification is assumed to be more accurate, and is simply the difference between the census and CPS estimates of the proportion of persons in a given category. For this study, the assumption of greater accuracy in CPS is not necessarily true for some characteristics. Therefore, the net difference rates do not always estimate error but merely differences in results obtained from the two data collection systems.

In general, the basic demographic and social characteristics (age, sex, race, household relationship, and marital status) exhibit a high level of response or classification consistency between the CPS and Census, as can be seen in table 1. For each, the estimated L-fold index of inconsistency was under 20 and, with the exception of the age classification for Negroes, the L-fold indices were under 10. These relatively low indices were observed for both men and women for each characteristic and for age classification by major residence categories (i.e. metropolitan, central city; metropolitan, outside central city; and nonmetropolitan, urban and rural). Also, none of the estimated indices differed appreciably from those ascertained in the 1960 CPS-Census matching study.

In table 2, the estimated indices of inconsistency and net difference rates are shown for each category in the distributions. Here again the age classification exhibits a high level of consistency as none of the indices for five-year age groups exceed 20 and most are under 10. Furthermore, the net difference rates indicate there are no substantial biases in the age classification.

Classification of the population by race is also highly consistent for whites and Negroes but slightly more inconsistent for persons of other races. Also, the net difference rates indicate a slight downward bias in the census classification of persons as white and a small upward bias toward classifying perions in races other than white or Negro.

Although the L-fold indices of inconsistency for household relationship and marital status are quite low, some specific categories within these distributions are somewhat more inconsistent. The relationship categories "other relative" and "nonrelative" have indices which approach or exceed 20. The classification "other relative" is one which would, most likely, be more accurately determined by an experienced CPS interviewer. However, the classification as a "nonrelative" could easily change from one enumeration to another. In households not occupied by members of the same family the person being interviewed or completing the questionnaire would be classified as a primary individual and all other persons in the household would be "nonrelatives." Obviously, the person answering the questions could be different between the CPS and census and the classification would, therefore, vary. With regard to the marital status classification, the categories "separated" and "divorced" were moderately inconsistent between the CPS and census. Finally, the net difference rates for both marital status and household relationship indicate that all categories are relatively free of bias.

Classification of persons by educational attainment was moderately inconsistent between the CPS and census. The L-fold indices are in the high thirties for both men and women. However, classification at the terminal levels of education, that is 4 years of high school, 4 years of college, and 5 or more years of college, was more consistent than was true for other attainment levels. Specifically, the indices for these categories were between 20 and 30. As is evidenced by the net difference rates, differences between the CPS and census classification by educational attainment are largely offsetting and, hence, little bias is noted. There is some indication, however, that the census may tend to understate slightly the terminal education categories relative to CPS. However, it may not be appropriate to view these differences as an indication of error in the census education statistics. There is some speculation that a respondent in a personal interview situation (as in CPS) may be more likely to erroneously report education at a terminal category than is true when the person is actually completing a questionnaire.

The estimated L-fold index of inconsistency for veteran status and period of service (15) reveals a low level of inconsistency between the two data sources. The reporting of veteran-nonveteran was relatively uniform as was the classification of World War II veterans. The remaining service categories, however, have index values near or above 20, and the estimated index for the residual "other service" category exceeds 30. As seen from the net difference rates, the census tends to understate the "other service" and "nonveteran" categories relative to CPS and to slightly overstate the proportion of veterans who served in the Vietnam and Korean conflicts.

The economic characteristics (i.e. employment status, work experience, class of worker, occupation, industry, and income) were, in general, less consistent in classification between the CPS and Census than was true for the demographic and social variables. This same relationship was observed in the 1960 matching study. In particular, the occupational classification for men and the distribution by weeks worked and most types of income for both men and women had estimated indices ranging from the low thirties to over fifty. On the other hand, the L-fold indices for employment status, class of worker, and industry with values of 20 or less indicate a relatively low level of inconsistency.

The indices of inconsistency for employment status cannot be strictly interpreted as measures of response or classification error. Since many persons could have experienced a change in employment status between the March 1970 CPS interview and the time of census enumeration, some portion of the difference in classification is valid. The indices reflect a combination of classification

errors and actual changes, and the index of inconsistency is, therefore, overstated to the extent that actual changes occurred. Even so, the L-fold index of about 20 for men and women indicates a reasonably high level of consistency. Among the four employment status categories, classification as "employed in nonagricultural industries" and as "not in the labor force" was fairly uniform between CPS and Census. However, the indices associated with the category "employed in agriculture" were somewhat higher, especially for women; and the classification "unemployed" was very inconsistent for both men and women. It must be remembered, though, that unemployment is subject to change over a short period of time and most of the response differences observed may reflect real changes.

The L-fold index of inconsistency on work experience for both men and women (43 and 37 respectively) reflects a moderately high level of disagreement between the CPS and census classification. However, the indices for specific categories indicate that the dichotomy of worked in 1969/did not work in 1969 and the identification of year-round workers (50 to 52 weeks) were fairly consistent. On the other hand, the identification of specific weeks worked categories for other than year-round workers was highly variable. The net result of differences in classification indicates that the census tended to understate the proportion working 50 to 52 weeks and to overstate the proportion who worked from 40 to 49 weeks and the proportion who did not work in 1969.

Of the three census "job content" classifications, class of worker and major industry group were reasonably consistent with the CPS. Among the seven class of worker categories, three had estimated indices under 20. However, the classification of unpaid family workers and of agricultural wage and salary workers was less uniform. All but four of the twelve major industry categories had estimated indices of inconsistency under 20 and none of the estimated indices exceeded 50. However, the net difference rates indicate a slight understatement of the proportion working in manufacturing industries in the census and a small overstatement for wholesale and retail trade.

The L-fold indices for major occupation (31 for men and 21 for women) reflect a moderate level of inconsistency in classification. For only three groups--farmers and farm managers, service workers, and private household workers--were the estimated indices below 20. In seven other groups the indices ranged from the low twenties to the low thirties. However, for two occupations--managers and administrators and laborers--the indices were about 40 and 50, respectively. Moreover, for managers the net difference rate reflects an understatement of about 3 percentage points in the census. However, evidence has indicated that the CPS occupation item, prior to a revision instituted in December 1971, tended to overestimate the proportion of managers. Since the comparisons made in this study are based on the old CPS occupation item, it may not be appropriate to view the census count of managers as negatively biased. For a number of occupation descriptions, the determination of major occupation group is very difficult. For instance, the distinction between a warehouseman (a laborer) and a fork-life operator (an operative) is often very subtle, based on the information at hand. As a result, there historically has been a great deal of inconsistency in classification especially among the "blue-collar" occupation groups. Since the laborer group constitutes a much smaller proportion of employed persons than do craftsmen and operatives, differences in classification among these groups have relatively more effect on the index for the smaller group.

Gross differences in income classification between the CPS and census resulted in a rather high level of inconsistency. The estimated value for the L-fold index of inconsistency for persons total income was 45 (50 for men and 43 for women). Among the type of earnings categories, wage or salary income was classified with a moderate level of inconsistency, as reflected by indices of inconsistency in the thirties. However, the classification by nonfarm and farm self-employment income was highly inconsistent. The type of "income other than earnings" which was most consistently reported was Social Security income. The limited range in the amount of income that can be collected under the Social Security system may explain the relatively low levels of inconsistency. The other sources of unearned income, including public assistance and "all other income," were characterized by high levels of inconsistency. The census category "all other income" was broken down into three separate questions in the CPS and the extra questionnaire detail may have helped the respondent in the CPS to recall small amounts of income from relatively unimportant sources.

The gross errors or differences in classification between the CPS and census seem to be largely offsetting since, at least for total income, there is little or no bias associated with the specific income categories. However, the proportion of persons reporting "no income" was overstated slightly in the census and the proportion reporting income under $1,000 was understated somewhat. This may again signify that respondents were more likely to recall small amounts of income in the CPS than they were in the census.

For the second portion of the study, a matching of Census and IRS data was accomplished by attempting to obtain 1969 Federal income tax returns for all individuals 14 years old and over in the CPS-Census sample. The 1970 Census is the third Census to be evaluated by using tax return data as an income benchmark[1]. The strict confidentiality of both the Census and IRS data make this type of matching evaluation difficult. All matching work was done by the Bureau of the Census in order to preserve the confidentiality of replies to census questions; no one other than Bureau employees, who are sworn to uphold the confidentiality of all Census information, had access to the information. The confidentiality of IRS records was also safeguarded. A detailed description of the procedures implemented to safeguard the confidentiality of the Census, CPS, and IRS data will be published by the Bureau in the final report on this project.

The Social Security Number (SSN) for persons 14 years old and over has been collected as part of the March Current Population Survey (CPS) for several years. These SSN's as well as name and address were used by Census Bureau employees to obtain tax returns. Once a tax return was located, the CPS identifying information was assigned to the IRS data and the name and SSN were not used in the actual match of the data.

For the cases the Census agents could not find, the majority were because no SSN was available from the CPS. For these additional efforts were made to obtain respondents' social security numbers from the Social Security Administration by name and birth date search of Social Security records. This procedure resulted in obtaining an additional 500 tax returns from the 2,000 needed bringing the final total to 8,434 returns located.

The assumption of the study is that Internal Revenue Service (IRS) income data is more accurate than Census income data. Since the definition of taxable income is different from the definition of total money income as used in the census, the comparison is not as straightforward as is often thought. In addition, not all persons are required to file tax returns, especially low-income persons and persons living on transfer payments. Consequently, the study is limited to persons filing returns located in the CPS or Census samples.

Six questions on income were asked in the 1970 Census. Listed below are the definitions of the income types from both sources. Wages and salaries is the only income type with the same definition in both Census and IRS. Other problems: (1) There was some tendency to report second job wages, in miscellaneous income (Schedule E) rather than wages and salaries, but this was limited to small amounts, (2) the use of Schedule F and C net income to approximate self-employed farm and nonfarm self-employment income is very rough because of the impossibility of splitting partnership income into farm and nonfarm sources, (3) the reporting of some sale of livestock as capital gains (Schedule D), and (4) the IRS rules defining current year expenses for farm income.

Census Income Type	IRS Income Type
Total Money Income (TMY)	Adjusted Gross Income (AGI)
1. Wages, salary, commissions, bonuses, and tips	Wages and Salaries
2. Earnings from nonfarm business, professional practice or partnership	a. Net income from Schedule C b. Partnership income from Schedule E 2/
3. Earnings from own farm	a. Net income from Schedule F 2/ 3/
4. Social Security or Railroad Retirement income	Not reported to IRS
5. Public Assistance or Welfare payments	Not reported to IRS
6. Other source - Interest, dividends, veterans payments pensions, and other regular payments	a. Schedule B - Gross Dividends and Interest 4/ b. Schedule E - Pension annuity net rent, net royalties, income from estates or trust, small business dividends, misc. income

The census data was tabulated for persons, family, and unrelated individuals. Tabulations from the matched Census and IRS file by persons is not possible because income on joint returns cannot be assigned separately to either the husband or the wife. The possibility of splitting wages and salaries reported on joint returns using W-2 Forms was considered, but the W-2 Forms were not available for many of the returns.

Although the data have been tabulated several ways, the data used in this paper are restricted to a matched household head, and household head and wife matched to either a joint return or separate returns. The data are preliminary.

As table 3 shows, the census total money income (TMY) was only 2.9 percent less than IRS's adjusted gross income (AGI). If the conceptual differences are taken into account as far as possible by removing capital gains from AGI and removing transfer payments (Social Security and public assistance) from Census total money income the Census aggregate is 3.4 percent less than the IRS aggregate. The consistency as measured by the L-fold index is high, 66.3 and 63.6 (see table 3).

Wages and salaries is the only type of income with an L-fold index below 50 when computed on the attached income intervals. Even when the classes are broadened to $5000 intervals, the inconsistency remains high at 33 (see table A). The Census captured 95.0 percent of the wages and salaries reported to IRS.

The L-fold indices for nonfarm and farm self-employment income were 69.8 and 77.0 percent respectively. The aggregate census farm income was over twice the net IRS farm income reported on Schedule F, but this is partly due to conceptual differences.

Table A.—NUMBER OF MATCHED HEADS OR MATCHED HEAD AND WIFE BY WAGES AND SALARIES INCOME IN THE 1970 CENSUS AND 1969 TAX RETURNS

IRS Wages or Salaries	Total	Under $5000	$5000 to $9999	$10,000 to $14,999	$15,000 and over
Total	6175	2128	2028	1309	709
Under $5,000	2043	1662	254	81	46
$5,000 to $9,999	1984	280	1513	157	34
$10,000 - $14,999	1395	131	200	980	84
$15,000 and over	752	55	61	91	545

L-fold index is equal to 33.2.

Summary

The small differences in Census and IRS aggregates for total money income, and wages and salaries are encouraging. However, the high inconsistency, even from wages and salaries, indicates there is still a lot of work to be done in improving the collection of all sources of income data.

FOOTNOTES

1/ The accuracy of income data collected in both the 1950 and 1960 Census was evaluated by using tax return data. The results of the 1960 effort were published in Record Check of Accuracy of Income Reporting, Series ER-60, No. 8, U.S. Bureau of the Census. The results of the 1950 study were published in An Appraisal of the 1950 Census Income Data, "Income Reported in the 1950 Census and on Income Tax Returns," Herman P. Miller and Leon R. Paley.

2/ Partnership income on Schedule E could be from farming but there is no way to separate partnership income by source.

3/ Some capital gains income (Schedule D) resulting from sale of livestock would also be considered farm income, but this cannot be easily identified.

4/ State tax refunds are supposed to be reported as interest, but interest identified as tax refunds was excluded from interest.

Table 3.—NUMBER OF MATCHED HEADS OR MATCHED HEAD AND WIFE BY TYPE OF INCOME FOR 1969 (PRELIMINARY)

Income Class Intervals	Total Income		Adjusted Total Income		Wages and Salaries		Nonfarm Self-Employment		Farm Self-Employment	
	Total Money Income (TMY)	Adjusted Gross Income (AGI)	TMY (Minus) (SS+PA) Transfers	AGI (Minus) Capital Gains	Wages and Salaries	Wages and Salaries	Nonfarm Self-Employment Income	IRS Net from Schedule C and Partnership income from Schedule E	Farm Self-Employment Income	IRS Net from Schedule F
	(Census)	(IRS)	(Census)	(IRS)	(Census)	(IRS)	(Census)	(IRS)	(Census)	(IRS)
Total Matched Units.....	6,174	6,174	6,174	6,174	6,174	6,174	6,174	6,174	6,174	6,174
LOSS	10	--	11	16	--	--	34	153	32	133
None.....	47	56	208	48	1,021	811	5,542	5,405	5,890	5,813
$1 - $999	124	170	191	180	172	254	86	140	65	80
$1,000 - $1,999	263	276	264	291	211	228	61	70	44	41
$2,000 - $2,999	286	318	275	325	200	208	41	52	31	28
$3,000 - $3,999	375	322	296	319	238	245	52	46	14	14
$4,000 - $4,999	359	372	341	372	286	297	33	28	13	12
$5,000 - $5,999	407	390	399	389	334	330	53	42	22	7
$6,000 - $6,999	481	392	441	390	436	379	38	30	12	14
$7,000 - $7,999	487	455	466	447	445	417	24	30	6	11
$8,000 - $8,999	500	491	489	499	431	431	22	26	15	6
$9,000 - $9,999	423	461	417	447	382	427	21	13	9	2
$10,000 - $11,999	742	733	725	734	702	682	34	23	12	7
$12,000 - $14,999	707	755	694	758	607	713	37	25	6	4
$15,000 - $24,999	747	781	744	770	597	649	40	44	2	1
$25,000 AND OVER	216	202	213	189	112	103	56	47	1	1
Median (dollars)1/.....	8,496	8,684	8,398	8,621	7,424	7,803	--	--	--	--
Aggregate income (dollars)...	60,066,846	61,882,002	58,578,912	60,653,376	48,027,546	50,552,712	5,291,118	4,568,760	913,752	407,484
Net percent difference2/.....	-2.9	(X)	-3.4	(X)	-5.0	(X)	+15.8	(X)	+124.2	(X)
L-fold Index of inconsistency.....	66.3		63.6		47.3		69.8		77.0	
95% confidence interval on index of inconsistency.....	65.0-67.7		62.3-64.9		45.7-48.9		65.0-75.8		70.4-86.1	

NA Not applicable. -- Equals zero.

1/ The none's were included in the computation of the median.

2/ $\frac{\text{Census-IRS}}{\text{IRS}} \times 100$ = Net percent difference. The 95 percent confidence intervals for the net percent difference are: Total income -5.5 to -0.3, adjusted total income -6.6 to -0.2, wages and salaries -6.0 to -4.0, nonfarm self-employment -4.5 to 36.1, and farm self-employment 37.1 to 211.3.