WHAT DO WE REALLY KNOW ABOUT WAGES:
THE IMPORTANCE OF NONREPORTING AND CENSUS IMPUTATION[1]

BY

LEE LILLARD, JAMES P. SMITH, AND FINIS WELCH

THE RAND CORPORATION

1700 MAIN STREET

SANTA MONICA, CALIFORNIA 90406

March 1982

INTRODUCTION

On April 1, 1980, 43 million Americans received the long form of the Census, which included questions covering many aspects of their lives, including their 1979 incomes. If experience with past Censuses and the Current Population Surveys is a useful guide, 7 to 9 million of these Americans will have refused to divulge their incomes. The Census Bureau has long been concerned with this problem and has been expanding its efforts, with limited success, to offset the increasing tendency for Americans not to respond to inquiries about their income. Apparently, the research community is less perturbed over the problem, judging by its use of the two major data files supplied by the Census, the 1-100 decennial U.S. Census and the yearly Current Population Surveys (CPS). In five of the leading economic journals, we counted over 100 articles during the last decade that used income data from these sources in their empirical analysis; not one of them made any effort to deal with the potential problems caused by nonrandom refusal to report income.[2] Perhaps one reason for this omission is that an income amount is coded for nonresponders, based on an income imputation procedure developed by the Census. Although the tapes include a flag warning that income was imputed, researchers have proceeded to treat the observations on income of responders and nonresponders as statistically equivalent.

This paper investigates some issues raised by the current Census treatment of nonresponses. In Sec. I, we briefly review the history of the Census Bureau's income imputation methodology. We also attempt to isolate characteristics that distinguish income reporters from

nonreporters and to assess the quality of the matches between nonincome reporters and "similar" reporters. Section II outlines the statistical techniques we used to account for nonrandom refusals to answer. Using recent statistical techniques for dealing with censored data, we contrast both instrumental variable and maximum likelihood procedures that explicitly incorporate the fact that some individuals will systematically refuse to report their incomes. Because we recognize that such estimates are sensitive to maintained distributional assumptions, we test this sensitivity by considering a series of Box-Cox transformations. Since one of the commonest uses of CPS and Census data involves the estimation of earnings functions, our empirical application focuses on estimating a standard earnings function for white males. The final section presents our statistical estimates and assesses the relative importance of nonresponse bias in empirical applications.

I. THE CENSUS TREATMENTS OF NON-RESPONSES

The Census bureau has continually revised its procedure for dealing with nonresponses. Until the early 1960s, individuals who refused to report their income were simply ignored in published data. Beginning with the 1960 Census and the 1962 CPS, the Census introduced its "hot deck" procedure to impute an income to nonreporters, by which a nonreporter was assigned the income of another individual to whom he was statistically matched for the variables included in the matching algorithm. Since its introduction, this "hot deck" technique has served as the basis for all subsequent income imputations, although the list of variables used, as well as the disaggregation permitted within each variable, has periodically been expanded. While such expansion seems desirable, the alteration in the income imputation algorithm itself raises questions concerning the comparability in Census and CPS data sets over time. For example, sex was not among the variables used to impute income until the 1968 CPS, and education was not added until the 1976 CPS! (See Appendix Table 23 for a brief history of the major changes in the CPS and Census income imputation procedure.)

The most fundamental recent change in methodology occurred between the 1975 and 1976 Current Population Surveys. Table 1 lists the characteristics used for the "new" (post-1975) and "old" (pre-1976) algorithms. The potential expansion in detail is indeed impressive, particularly for occupation, education, age, and residential location. However, we will show below that this potential is frequently not realized. Beginning in 1976,[3] in addition to expanding the list of

Table 1

COMPARISON OF CHARACTERISTICS USED TO IMPUTE MISSING EARNINGS RESPONSES: "OLD" VS. "NEW" PROCESSING PROCEDURES[a]

Old	New
Occupation of Longest Job	
Professional Managerial Sales Clerical Crafts Operatives Service Nonfarm laborers Farm laborers Farmers and farm managers	the 3-digit occupation classification
Age	
14-24 years 25-44 years 45-64 years 65 years or more	14-17 years 18-24 years 25-34 years 35-44 years 45-54 years 55-64 years 65 years or more
Family Relationship	
Head Other family member (including wife) or unrelated individual	Head Wife Other family member Unrelated individual
Region	
Not used	South Non-South

Old	New
Weeks Worked and Full-time/Part-time Status	
50-52 weeks, full time 14-49 weeks, full or part time 1-13 weeks, full or part time	50-52 weeks, full time 48-49 weeks, full time 40-47 weeks, full time 27-39 weeks, full time 14-26 weeks, full time 1-13 weeks, full time (The above 6 items, part time)
Race	
White Negro and other	White and other, excluding Spanish Negro, including Negro Spanish White Spanish
Residence	
Not used	In SMSA, size 1,000,000+ In SMSA, size under 1,000,000 Not in SMSA
Sex	
Male Female	Male Female

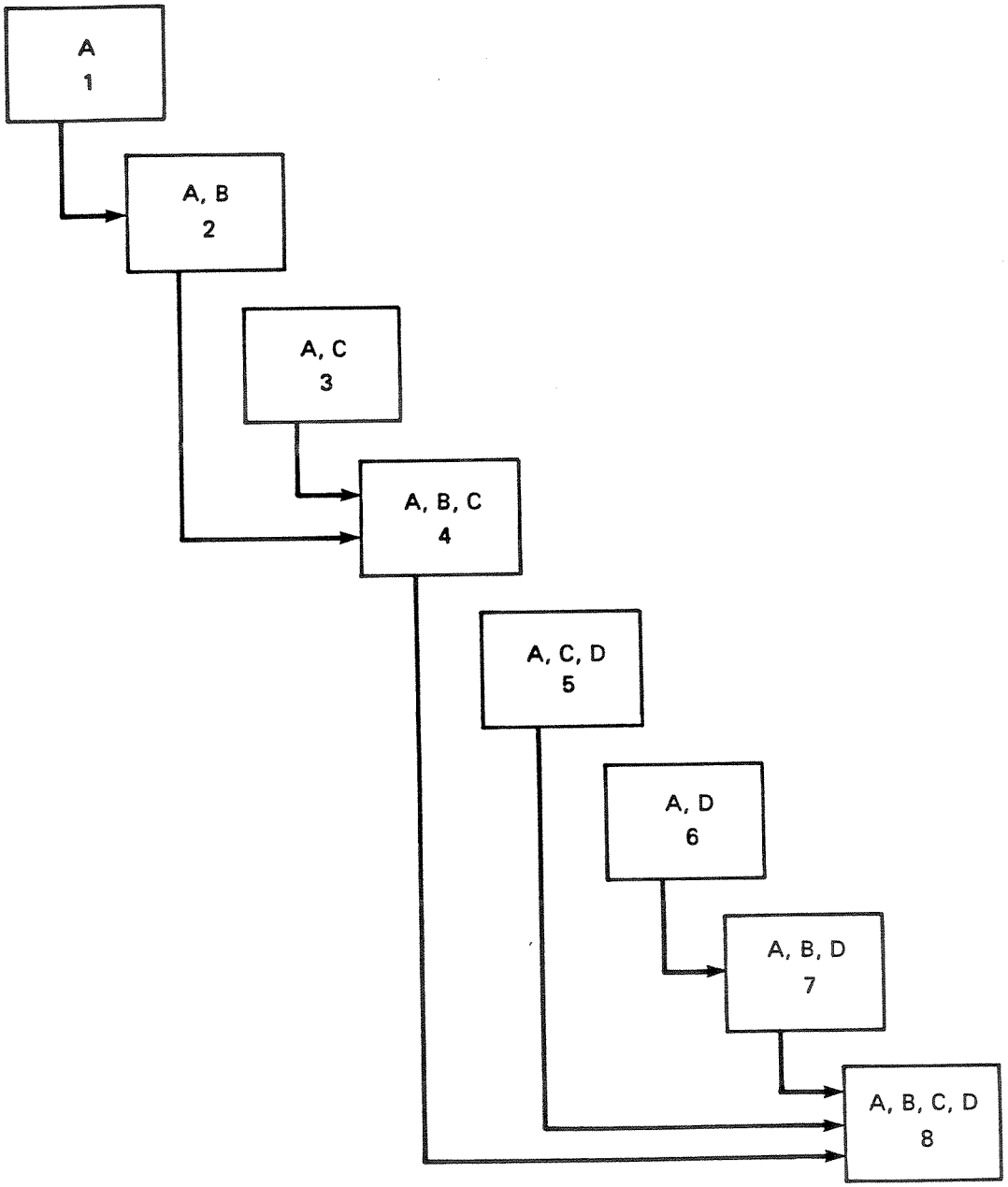
Old	New
Class of Worker	
Wage or salary Self-employed Without pay in family business	Private wage or salary Federal government wage or salary State or local government wage or salary Self-employed Without pay in family business
Educational Attainment	
Not used	0-11 years 12 years 13-15 years 16 years 17 or more years
Labor Force Status of Spouse	
Not used	Spouse in labor force Spouse not in labor force

[a] The characteristics and detail within characteristics listed here represent the maximum detail used in the imputation procedure. While all characteristics listed for the "New" system would be used for any person requiring an imputation of missing earnings response(s), the detail listed for the "Old" system is not used for every such person, but varies considerably by sex, race, and age group. This table is contained in a Census Bureau publication, "Revisions to the March CPS Processing System."

characteristics, the Census Bureau simultaneously matched nonrespondents to questions on earnings, work experience (weeks worked, full-time/part-time status), and longest job held (occupation and industry). A nonrespondent on all three questions would receive his imputed values for all variables from the same respondent. Depending on an individual's combination of missing values for earnings,[4] work experience, and longest job, a civilian[5] nonrespondent is initially placed into one of eight groups. Given the initial placement group, the Census searches to obtain a match with a similar respondent, using a variable-list key consisting of a subset of the variables and within-variable disaggregation displayed in Table 1. Moreover, within each group there are up to four layers of aggregation on the variables used. Therefore, the Census first attempts to match each nonrespondent with a similar respondent at the lowest level of aggregation possible in his group. If no respondent match is found in the initial assignment group, the nonrespondent is then moved into a group that generally tends to be less restrictive on the qualifications for a match. Figure 1 depicts the criteria that determine initial assignment as well as the group reassignment path if no match is discovered in a group. Since the final level of aggregation in group 8 utilizes only sex and three broad age and education groups, all nonrespondents are eventually matched.

Who are the Nonreporters?

Table 2 lists the percentages of white and black males who did not report incomes for each of the individual March 1968-1978 CPSs. These percentages vary considerably from year to year, but they are not



A - Earnings amount missing
B - Recipiency missing
C - Work experience missing (e.g., weeks worked, etc.)
D - Longest job missing (e.g., occupation, etc.)

Fig. 1 — Reclassification pattern to obtain matches

Table 2

PERCENTAGE OF SAMPLE WITH FLAG
FOR INCOME IMPUTATION

Group	Year of Current Population Survey										
	1968	1969	1970	1971	1972	1973	1974	1975	1976[a]	1977	1978
White males	11.5	14.0	10.7	10.8	10.5	12.0	14.2	15.7	17.3	14.7	15.7
Black males	14.5	14.6	10.3	11.3	10.7	12.6	15.0	14.9	18.0	14.6	15.9

[a] Beginning in 1976, the Census flag indicated that income was imputed for a particular individual. In prior years, the flag indicated only that income was imputed for some individual in the family. We constructed a family flag for the 1976-78 CPSs, and found that the percentages flagged were 22.4, 27.3, 21.0 for white males and 23.3, 29.0, 21.9 for black males.

trivial and the problem appears to have become more serious over time.[6] Through 1974, nonreporters ranged from 10 to 15 percent but usually exceeded 15 percent thereafter.

It is well established that nonrespondents have personal characteristics that will assign them higher incomes than the average respondent. To illustrate, Table 3 lists the ratio of mean Census-imputed income of nonrespondents relative to mean income of respondents. Imputed incomes among nonrespondents exceed those of respondents by 2 to 10 percent. Given the income imputation procedure used by the Census, this necessarily implies that nonrespondents have higher levels of attributes that produce higher income than respondents do. We can also detect from Table 3 one consequence of the expanded set of variables used beginning in 1976. The enlarged within-variable detail, particularly for occupation and education, results in a much larger difference in the mean income of the two groups.[7]

Table 4 provides a more detailed look at the relation between income and reporting status. White males in the 1980 CPS survey were placed into one of a series of earnings intervals based either on their self-reported earnings (for respondents) or their Census-imputed earnings (for nonrespondents). In each interval, we list the percentage of men who did not report their earnings. The simple relation between reporting status and income is non-monotonic. The probability of not reporting is substantially larger for earnings above $40,000, where the figure approaches 30 percent. However, over the broad range of positive earnings up to $30,000, the relationship is U-shaped--hitting its trough in the $16 to $19 thousand interval. Other factors, in addition to

Table 3

RATIO OF CENSUS-IMPUTED INCOME TO INCOME OF RESPONDENTS

	Years				
Group	68-69	70-71	72-73	74-75	76-78
White males	0.99	1.04	1.06	1.02	1.10
Black males	1.00	1.03	1.03	1.00	1.10

NOTE: Income is in 1968 dollars.

Table 4

PROPORTION OF NONREPORTING WHITE MALES, BY EARNINGS INTERVAL

Earnings Interval ($)	Wage, Self-Employment and Farm Earnings	Wage and Salary Earnings
1- 2999	19.8	18.7
3000- 5999	19.2	17.3
6000- 8999	18.1	16.4
9000-11999	17.4	14.9
12000-15999	16.6	14.2
16000-18999	15.3	14.0
19000-24999	16.9	15.5
25000-29999	18.2	17.1
30000-34999	18.5	16.1
35000-39999	20.5	19.3
40000-49999	27.0	23.5
50000+	28.1	26.5

NOTE: Data are based on 1980 CPS, and earnings refer to 1979.

income, clearly affect these reporting propensities. Some of these factors tend to be more prevalent among those men with the lowest incomes.

Table 5 lists the proportion of nonreporters stratified by occupation, type of income, and rotation month. The table isolates those occupations in which nonreporting is significantly higher than the norm. These occupations have at least one of the following two characteristics in common: They are among the highest-income occupations, or considerable ambiguity clouds the calculation of net income from receipts and expenses for income tax purposes. Lawyers, doctors, and dentists fit both criteria, and fully one-third do not answer income questions. Similar proportions are listed among the farmers and private household groups, where incomes typically are much lower, but where income tax considerations loom as large.

Table 5 lists nonreporting proportions by type of earnings. Fully 16 percent of men with wage and salary earnings refuse to report the amount of their income. Not surprisingly, refusals are substantially higher among those with self-employment income--either on the farm or not--with one in four men with this type of earnings listed as nonreporters.[8]

The sampling design of the CPS surveys may indirectly affect the aggregate income reporting probabilities. Each respondent is scheduled for eight monthly interviews, conducted in two stages. After initial selection, a respondent is interviewed for four successive months, exits the sample for eight months, but reappears once again for four more monthly surveys. Clearly, not all respondents elect to remain for the

Table 5

PROPORTION OF NONREPORTING WHITE MALES, BY OCCUPATION STATUS, TYPE OF EARNINGS, AND ROTATION MONTH

A. By Occupation Status

Occupation	Proportion
Lawyers and judges	28.6
Dentists	33.6
Doctors	33.0
Engineers	17.7
Other professionals	14.0
Managers	20.6
Farmers	25.3
Private households	26.3
All others	15.7

B. By Type of Earnings

Class of Worker	Proportion
Wage and salary	16.3
Self-employed nonfarm	25.7
Self-employed farm	24.7

C. By Rotation Month

Month	Proportion	Mean Income	Unemployment Rate
1	15.8	15316	6.49
2	18.6	15516	5.78
3	18.9	15642	5.61
4	17.9	15512	5.36
5	19.5	15606	5.38
6	19.5	15408	5.81
7	19.8	15505	5.32
8	18.4	15745	5.31

duration. If those who remain in the sample from month to month differ systematically from those who decide to drop out, nonreporting proportions will not be independent of rotation month. In addition, the first interview in each stage (months 1 to 5) is a personal interview while the others are conducted by telephone. Personal interviews presumably are more effective than telephone surveys in eliciting answers to income questions.

A crude check on these notions is provided at the bottom of Table 5. The percentage of nonreporting white males is decidedly lower in the first month (a personal interview) and appears somewhat larger during the second cluster of four months than the first.[9] Mean income, and particularly unemployment rates by rotation month, suggest that the probability of remaining in the CPS sample is not random. Between months 1 and 8, mean income rises by 2.8 percent while the unemployment rate falls by almost 20 percent. CPS dropouts tend to be more concentrated among those with low income and the unemployed. The reporting percentages in Table 5 do appear to mirror the mean income changes across rotation months, although the range of differences is not overwhelming.[10]

Table 6 illustrates the type of problems that can arise when the imputation algorithm is changed. During the 1968-1975 period, the ratio of imputed to non-imputed incomes declines sharply with years of schooling. By not using schooling, we overstate the income of nonreporters who have little schooling and understate incomes of more highly educated nonreporters. However, this pattern is eliminated over the 1976-1978 data when education is added to the list of variables

Table 6

RATIO OF IMPUTED TO NON-IMPUTED ARITHMETIC MEANS OF WAGE AND SALARY INCOME, AND PROPORTION OF SAMPLE FLAGGED, BY EDUCATION LEVEL

A. Ratio of Imputed to Non-imputed Means

	Years of Schooling, 1968-1975[a]			Years of Schooling, 1976-1978[a]		
Sample	0-8	9-12	13+	0-8	9-12	13+
White males	1.17	1.04	.950	1.13	1.04	1.13
Black males	1.18	0.985	.880	1.19	1.04	1.11
White females	1.22	1.10	.935	1.03	1.02	1.01
Black females	1.31	1.08	.800	1.12	1.23	1.08

B. Proportion of Sample Flagged

	Years of Schooling Completed[b]		
Sample	0-8	9-12	13+
White males	12.4	13.1	14.3
Black males	13.0	14.1	14.1

NOTE: Wage and salary income is in real 1968 dollars.

[a]Data are pooled across indicated years.

[b]Data are pooled across all 11 years of CPS data.

utilized to match a nonrespondent to a "similar" respondent. Indeed, over these last three years, imputed incomes of nonreporting college graduates exceed those of college graduate reporters, perhaps reflecting the added occupational detail used after 1975.

How Accurate are Census Matches?

The Census income imputation algorithm is similar to a fully interactive analysis of variance where an empirically based estimate of residual variance is implicitly added back to the income assigned to nonreporters.[11] The qualification is that the set of indicators employed varies over nonrespondents, depending on initial assignment or on how thin the sample is on their potential matches. For example, blacks would typically be matched at much higher levels of aggregation than whites, and women at higher levels than men.

Deferring to Sec. II the issue of nonrandom refusals to report, the Census method of handling missing income values has the advantage of remaining relatively agnostic on the appropriate functional form of the determinants of earnings or on the distribution of residual variances. The most straightforward alternative would be to posit and estimate over reporters an explicit model of income determination. With these regression estimates, one could impute an expected income to each nonreporter, which would be augmented by a random draw from a normal distribution, with a variance estimate obtained from the regression on reporters.

If matches could be found for all nonreporters using something close to the detail depicted in Table 1 under the new list, we would

have little quarrel with the Census technique.[12] We are less sanguine if, for many individuals or within important subsets of the population, initial matches are not found and we are forced to proceed to Census fall-back positions. The desirability of tying the accuracy of a match to the likelihood of uncovering a similar person is not at all evident. Moreover, when the Census decides to drop a control variable from its assigned list, such as region of residence, it is forced to ignore not only the interactive effects of region with other characteristics, but the direct effect of region itself. The alternative of specifying an explicit model offers more flexibility in choosing the appropriate fall-back position.

To get some notion of the extent of these problems, Table 7 presents the distribution of nonreporters of income by their initial level of assignment group (defined in Fig. 1) as well as the group in which a match was eventually obtained. Initial assignment depends also on the missing variables, while final assignment group level depends on the ability to find a match. For initial assignment, nonreporters are predominately placed into two groups: (1) those who did not report the amount of earnings, but did report work experience and longest job (group 1), and (2) those who reported neither earnings nor work experience, but did report longest job (group 4). A comparison of the initial and final group illustrates the effects of sample thinness on the probability of a successful match. The proportion of women and particularly blacks who were moved down the line to obtain matches is significantly larger than it is for white men. For example, one-third of all relevant blacks could not find a match in group 1, but only about 10 percent of the white men.

Table 7

DISTRIBUTION OF NONREPORTERS BY CENSUS ASSIGNMENT GROUP

Initial Assignment Group	Criteria for Inclusion in Group[a]	White Men	Black Men	White Women	Black Women
1	A	46.8	40.8	41.3	33.7
2	A,B	7.7	5.7	7.1	4.8
3	A,C	0.7	1.0	0.7	1.5
4	A,B,C	40.6	46.2	42.3	50.5
5	A,C,D	0.0	0.1	0.1	0.0
6	A,D	0.2	0.4	0.3	0.5
7	A,B,D	0.3	1.1	0.6	1.1
8	A,B,D,C	3.7	4.6	7.7	8.0
Final Group					
1	A	42.5	26.8	36.1	21.7
2	A,B	8.7	4.6	7.3	3.8
3	A,C	0.6	0.7	0.6	1.1
4	A,B,C	43.1	52.9	45.6	57.3
5	A,C,D	0.0	0.1	0.4	0.0
6	A,D	0.2	0.4	0.3	0.5
7	A,B,D	0.3	1.1	0.6	0.9
8	A,B,C,D	4.7	13.3	9.6	14.9

[a] The letters in this column refer to the following:

A - Earnings Amount Missing;

B - Recipiency Missing;

C - Work Experience Missing (e.g., weeks worked, etc.);

D - Longest Job Missing (e.g., occupation, etc.).

The large number of individuals who were initially placed into group 4 demonstrates that a significant number of individuals who did not report income also gave no information on work experience. The Census fall-back procedure when no initial match was found (see Fig. 1) contains an unfortunate error which compounds the extent of multiple missing values. When a nonrespondent is moved into another group, he is treated equivalently in all respects to a nonrespondent who began in the fall-back group.

For example, consider those who started their imputation lives in group 1 or 2 (i.e., did not report earnings, but did report work experience and longest job), but who were moved to group 4 because no match was found. Although our hypothetical nonrespondents did report their work experience, their reported values would be overridden and would be substituted by data from the matched donor in group 4. Table 8 lists the proportion of valid work experience and longest job information that was overridden. Because it appears first in the eight-group hierarchy of Fig. 1, the frequency of overriding is considerably more common for work experience than for longest job: 14 percent of otherwise valid work experience and 2 percent of longest job information was eradicated. This problem is clearly more serious when finding an initial match proves difficult. One-third of all relevant blacks have valid work experience answers changed to matched values. The comparable figure for black occupation and industry data runs between 7 and 9 percent. Even among white men, the Census eraser will be busy in the thin segments of the white sample. Combining those who did not report with those who had data overridden, Table 9 lists the proportion of

Table 8

PERCENTAGES OF OVERRIDDEN DATA
FOR PEOPLE REPORTING WORK EXPERIENCE
OR LONGEST JOB, BUT NO EARNINGS

Sample	Work Experience Overridden	Longest Job Overridden
White males	6.2	1.0
Black males	31.4	9.2
White women	10.2	2.2
Black women	33.1	7.5
All	13.9	2.1

income nonreporters who had either work experience or longest job allocated. More than half of all men with imputed income also had imputed work experience, while 8 percent of them had imputed occupation and industry.

Nonreporting of items other than income provides a partial explanation for the U-shaped relationship between Census income and income reporting status depicted in Table 3. Table 10 lists the percentages of white males who failed to report different combinations of income, work experience, and longest job data. This table suggests that it is useful to distinguish between two types of nonreporters. The first, general nonreporters, refuse to answer a wide array of questions besides the one on income. The second type, specific nonreporters, single out income as a sensitive item that they will not divulge. General nonreporters--those failing to report only items other than income, and those failing to report both income and other items--are actually most frequent in the bottom tail of the income distribution in Table 10 (i.e., below $6000); they probably have only sporadic and irregular contact with the labor market and are generally reluctant to answer any survey questions at all--which may well account for this group's large fraction of refusals to report income. In contrast, specific nonreporters are differentially singling out income as an item they will not answer. This selective sensitivity is monotonically related to income level. The last column of Table 10 lists the fraction of all nonreporters who refuse to report only their incomes. This ratio starts at 39 percent for those with Census incomes below $3,000 and increases across each income interval until it reaches a peak of 70

Table 9

PROPORTION OF NONREPORTERS
OF INCOME WHO HAD WORK EXPERIENCE
OR LONGEST JOB ALLOCATED

Sample	Work Experience Allocated	Longest Job Allocated
White men	48.4	5.2
Black men	66.9	14.9
White women	56.2	10.9
Black women	60.0	16.3
All	53.0	8.1

Table 10

PROPORTION OF NONREPORTERS BY ITEM NOT REPORTED

Income Interval	(1) Any Item	(2) Earnings Only	Earnings and Others	Others Only	Column 2/ Column 1
1 - 2999	21.3	8.4	9.8	3.1	39.4
3 - 5999	18.8	7.9	9.2	1.6	42.2
6 - 11999	16.5	7.6	8.1	0.5	46.1
12 - 24999	15.1	8.3	6.6	0.2	54.9
25 - 34999	17.0	9.9	7.3	0.1	58.2
35 - 39999	20.0	12.5	7.1	0.4	62.6
40 - 49999	24.3	15.1	8.6	0.7	62.1
50 +	26.0	18.1	7.9	0.0	69.6

NOTE: Sample consists of white males on 1980 CPS. "Other" items are work experience and longest job.

percent in the open-ended interval of $50,000 and above. We view the last column of Table 10 as strong evidence that income levels do affect the propensity to report income.

In light of the data presented in this section, how accurate are the Census matches, particularly compared to the precision implied by Table 1? We address this issue first in Table 11, which lists the fraction of matches in which the Census explicitly uses a characteristic in order to obtain a match.[13] It is clear that Census matches fall well short of completely using all the variables listed in Table 1. Only one in five white men are made to match in the four Census geographic regions, and slightly less than two-thirds are made to match on a North-South dichotomy. As we now should expect, the situation is more serious for black males. For them, less than one in ten matches use the four Census regions and less than half rely at least on a North-South division. Given the still substantial North-South earnings differentials that exist for blacks, ignoring this information seems particularly inadvisable. A similar story emerges for most of the other characteristics listed in Table 11. In less than one in five matches are the variables for marital status or number of children employed. It is somewhat more encouraging that two-thirds of all white matches use data on wives' labor force status, while slightly less than half of black matches use this characteristic.[14]

In only one-third of all male matches are the 3-digit occupation codes used, with very large differences once again existing between the races. Matches are predominately made instead at the fallback 46 occupation groups. Table 12 indicates that, at least for some

Table 11

PROPORTION OF NONREPORTING MATCHES IN WHICH INFORMATION IS USED

Characteristic	White Men	Black Men
Region		
4 Census regions	21.2	8.8
South/non-South only	42.4	36.6
SMSA ≥ 1,000,000 and farm status	5.3	12.4
SMSA ≥ 1,000,000 only	76.7	45.3
Occupation		
3-digit codes	37.1	9.2
46 occupation groups	57.7	75.8
Weeks worked, full-time/part-time	51.6	33.1
Labor force status of wife	65.0	47.2
Marital status	20.5	16.3
Number of children	19.8	12.3

occupations, the income assigned to a nonreporter can differ substantially depending on whether one uses an exact 3-digit occupation match or the 46 occupation groupings. In Table 12, we report mean earnings that the Census assigned to nonreporting lawyers and judges, stratified by level of occupation match. Judges with exact matches earned almost $30,000 more than those with non-exact matches. A similar comparison for lawyers yields a difference of almost $20,000. There is little question that the Census algorithm results in a substantial understatement of the income assigned to nonreporters in these two professions.[15]

Table 13 provides a more complete perspective on the quality of matches. The first part of the table compares mean differences in characteristics of nonreporters with their matched donors. The second part measures the standard deviation of these differences. In terms of age, education, hours, or weeks worked, nonreporters and their donors are almost identical on average.[16] More revealing, however, are the measures of spread.[17] Speaking loosely, a 10-year age difference, 3-1/3 years of schooling, and 7 weeks of work will cover two-thirds of all matches. Such differences do not seem trivial and they are even larger among black men. We note as well that the accuracy of matches on weeks worked and schooling are quite poor in the lower income intervals.

The evidence we presented in this section suggests that it may be time for the Census to rethink its imputation methodology. Most obviously, the overriding of data is both unnecessary and inappropriate. But more basically, the indirect evidence we have presented suggests

Table 12

MEAN INCOME OF NON-REPORTING JUDGES
AND LAWYERS BY OCCUPATION MATCH OF DONOR

Sample	Exact 3-Digit Match	Match with 46 Occupation Groups
Judges	$47,500	$18,223
Lawyers	33,448	15,594

NOTE: Sample consists of 1980 CPS white males.

Table 13

SUMMARY STATISTICS COMPARING NONRESPONDENTS WITH THEIR STATISTICAL MATCHES

Sample	Mean Difference (Reporters - Non-reporters)				Standard Deviation of Differences				Fraction Who Match Exactly at		
	Age	Ed.	Hours[a]	Weeks[a]	Age	Ed.	Hours	Weeks	3-Digit Code[a]	Census Region	South/ Non-South
White males	-.29	.01	.07	-.08	5.08	1.68	10.9	3.59	.556	.442	.850
Black males	.05	.10	.26	-.14	5.87	2.15	8.1	5.18	.374	.539	.759
White male income ($)											
<3000	-.20	.06	.14	-1.26	4.29	1.84	11.1	7.03	.593	.482	.860
3000- 6000	-.25	-.02	1.52	.46	4.76	2.02	12.3	5.55	.503	.454	.817
6000-12000	-.48	.01	.03	.22	5.41	2.10	10.1	3.13	.521	.441	.831
12000-25000	-.28	-.06	-.22	.06	5.18	1.50	10.4	1.58	.542	.426	.863
25000-35000	-.40	.06	-.38	.09	5.42	1.13	10.8	1.39	.590	.427	.846
35000-39000	-.01	.19	.85	-.05	5.18	.44	13.1	.51	.630	.454	.887
40000-49000	.10	.11	1.12	-.10	5.27	.33	10.3	.50	.608	.403	.818
50000+	-.10	.10	-.22	.01	5.27	.31	14.8	.56	.738	.415	.817

[a] Applies only to cases in which data were not allocated.

that statistical models of earnings functions seem likely to outperform the Census imputation algorithm. The basic problem is that the Census technique is linked not to the information content provided by variables in explaining earnings, but rather to that provided by the commonality of variables within the population. In those parts of the CPS samples that are thin, we have little doubt that a reasonably parsimonious earnings function would be superior to the Census method. In particular, black males certainly could be more accurately handled by an estimated earnings function over the sample of reporters.

II. A MODEL OF NONRANDOM NONREPORTING

The key assumption of the Census imputation procedure--that the probability of reporting earnings is random--is easily challenged. The actions of nonreporters suggest that they distinguish (even if the Census does not) between personal attributes that are correlates of income and income itself. Respondents were willing to answer questions about most of the characteristics that the Census uses to predict their incomes. It was their income they refused to divulge. The effect of earnings on the reporting propensity could reflect a combination of factors: an income-elastic demand for privacy, a fear of governmental or other uses of the data (particularly for income tax purposes presumably correlated with marginal tax rates), or simply a higher price of time for completing the survey. Thus, we expect higher earnings to reduce reporting propensities.

If the decision to report earnings to the CPS is nonrandom, reporters will systematically differ from nonreporters. The statistical procedures developed here--a variant of the familiar sample selection model--represent an attempt to quantify this relationship and any biases which might arise from it. This section describes both instrumental variable and maximum likelihood estimators for our two behavioral equations, the determinants of market earnings and the propensity to report those earnings in surveys.[18] Because we are concerned both with the possibility that income response is not random (as assumed in the Census algorithm) and the change in the list of variables used to impute income, we use four data sets in our empirical work: the 1970, 1975,

1976, and 1980 CPSs. With these data files, we estimate an earnings function for employed white males. We defer to Sec. III our discussion of the list of regressors we use in each of the behavioral equations.

A difficulty with recent sample selection models is the sensitivity of the resultant estimates to untested distributional assumptions. In particular, departures from the conventional normality assumption can easily confound our ability to retrieve meaningful estimates of the impact of sample selection bias. Although we maintain the normality assumption throughout, we do offer a partial generalization by considering the class of power transformations of earnings which include the now standard log transformation and dollar earnings as special cases. The appropriate transformation is to be estimated.

More formally, we assume the existence of an earnings (Y) function of the form

$$(2.1) \qquad (Y^{\theta} - 1)/\theta = \beta' X + u \ ,$$

where u is normally distributed with zero mean and standard deviation σ_u. This is the well-known Box-Cox (1964) power transformation.[19] For brevity, we use the notation $Y^* = (Y^{\theta} - 1)/\theta$. Note that for $\theta = 1$, the earnings function is in dollar terms, and as θ approaches zero the earnings function takes the log-linear form.[20] It will be useful later to note that for the marginal density of Y^*, $f(Y^*)$, and Y, $g(Y)$,

$$(2.2) \qquad f(Y^*)dY^* = \frac{1}{\sigma_u}\,\phi\,\left(\frac{Y^*-\beta'X}{\sigma_u}\right)dY^* = \frac{Y^{\theta-1}}{\sigma_u}\,\phi\,\left(\frac{Y^{\theta-1}-\beta'X}{\sigma_u}\right)dY = g(Y)dY \ .$$

The propensity to report earnings (Z) is given by

$$Z = \alpha' X + \gamma Y^* + v \tag{2.3a}$$

or in reduced form,

$$\begin{aligned} Z &= \alpha' X + \gamma\beta' X + e \\ &= Z(X) + e \ , \end{aligned} \tag{2.3b}$$

where

$$e = v + \gamma u \ .$$

Earnings are reported if $Z > 0$ and not reported otherwise. Define the indicator function

$$I = \begin{cases} 1 & \text{if } Z > 0 \\ 0 & \text{if } Z \leq 0 \end{cases} .$$

The propensity to report earnings may covary with earnings for two reasons: (1) The propensity to report is a function of earnings ($\gamma > 0$); and (2) the residuals in the earnings function and the propensity to report equation covary (censoring bias, $E(u \mid v) \neq 0$). Both reasons are incorporated in Eqs. (2.1)-(2.3).

Parameters of the model include the regression coefficients α, β, and γ and the elements of the covariance matrix Σ_{uv} subject to an arbitrary normalization of the scale Z.[21]

The normalization is

$$(2.4) \qquad \sigma_{ee} = \sigma_{vv} + 2\gamma\sigma_{vu} + \gamma^2\sigma_{uu} \equiv 1$$

so that

$$(2.5) \qquad \Sigma_{eu} = \begin{bmatrix} 1 & \gamma \\ 0 & 1 \end{bmatrix} \Sigma_{vu} \begin{bmatrix} 1 & 0 \\ \gamma & 1 \end{bmatrix} = \begin{bmatrix} 1 & \\ \sigma_{vu} + \gamma\sigma_{uu} & \sigma_{uu} \end{bmatrix} .$$

We assume that u and v are jointly normally distributed.

The Census imputation procedure effectively replaces the unobserved earnings value with a random draw from among observations within the sample with the same X values. The random draw is inappropriate when $\sigma_{eu} \neq 0$, i.e., when either $\sigma_{vu} \neq 0$ or $\gamma \neq 0$. The expected value of the random draw is not $\beta'X$.

$$(2.6) \qquad E(Y^* | X, I=0) = \beta'X + \sigma_{eu}\,\lambda_n(X)$$

where

$$(2.7) \qquad \lambda_n(X) = E(e | e \leq -Z(X)) = -\phi(-Z(X))/\Phi(-Z(X))$$

and ϕ and Φ are the standard normal density and distribution functions, respectively. Similarly, a regression of Y^* on X is inappropriate for reporters.

$$E(Y^* | X, I=1) = \beta' X + \sigma_{eu} \lambda_r(X) \tag{2.8}$$

where

$$\lambda_r(X) = E(e | e > -Z(X)) = \phi(-Z(X))/\Phi(Z(X)). \tag{2.9}$$

Parameters of the model are estimable by either instrumental variable or maximum likelihood methods, assuming at least one zero restriction on α. The reduced form parameters $(\alpha+\gamma\beta)$ are estimable using a simple probit model. The probit likelihood is given by

$$\mathcal{L}((\alpha+\gamma\beta) | X_i, I_i) = \prod_{i=1}^{N} \Phi((2I_i-1)(\alpha+\gamma\beta)' X_i) \tag{2.10a}$$

An alternative notation is more convenient later.

$$\mathcal{L} = \prod_{i=1}^{N} L_i \tag{2.10b}$$

where

$$L_i = \begin{cases} \Phi(-(\alpha+\gamma\beta)' X_i) & \text{if } I_i = 0 \\ \Phi((\alpha+\gamma\beta)' X_i) & \text{if } I_i = 1 \end{cases}. \tag{2.10c}$$

With consistent estimates of $(\alpha+\gamma\beta)$ we can construct instruments for $Z(X)$ and thus $\lambda_r(X) = \phi(-Z(X))/\Phi(Z(X))$. OLS may be applied to Eq. (2.8) to obtain consistent estimates of β and σ_{eu}. The earnings equation error variance σ_{uu} may be estimated from the OLS sum of squared residuals, SSE:

$$\hat{\sigma}_{uu} = SSE/N_r - \hat{\sigma}_{eu}^2 \mathcal{W} \tag{2.11}$$

where

$$\mathcal{W} = \sum_{i=1}^{N_r} (\hat{Z}(X_i)\hat{\lambda}_r(X_i) - \hat{\lambda}_r^2(X_i))/N_r \tag{2.12}$$

and N_r is the number of workers who report earnings. If there is at least one j such that $\alpha_j = 0$, and $\beta_j \neq 0$ then γ may be estimated as

$$\hat{\gamma} = \sum_{j=1}^{NJ} (\hat{\gamma}\beta_j/\hat{\beta}_j)/NJ \ . \tag{2.13}$$

This is the average of the multiple estimates if $NJ > 1$. Any one of the estimates or any other linear combination may be chosen. Alternatively, the multiple estimates for γ may be resolved to a single estimate by a second-stage probit where $Y^*(X_i) = \beta' X_i$ is used, $Z = \alpha' X_i + \gamma Y^*(X_i) + e$.

This clearly illustrates that the estimate of γ depends on a proportionality restriction on the effect of variables with a priori zero restricted coefficients.

Let us introduce an additional problem which is not adequately handled by the instrumental variable procedure. The CPS data report earnings as \$50,000 for all persons whose earnings are at or above that level. This presents an open-ended interval for the upper tail. To implement the instrumental variable procedure, we have assigned an earnings value of \$75,000 to those workers. The maximum likelihood procedure deals with the problem directly and is developed to allow estimation of θ as well.

The likelihood function is given by

$$
(2.14) \qquad \mathcal{L}_i = \begin{cases} \Phi(-(\alpha+\gamma\beta)'X_i) & \text{if} \quad I=0 \\ (Y^{\theta-1}/\sigma_u)\phi((Y^*-\beta'X_i)/\sigma_u) & \\ \quad X\ \Phi(((\alpha+\gamma\beta)'X_i+(Y_i^*-\beta'X_i)/\sigma_u)/(1-\rho_{eu}^2)^{1/2}) & \\ \qquad \text{if} \quad I=1 \quad \text{and} \quad Y < \$50{,}000. & \\ \Phi((\alpha+\gamma\beta)'X_i,-(Y50^*-\beta'X_i)/\sigma_u \mid \rho_{eu}) & \\ \qquad \text{if} \quad I=1 \quad \text{and} \quad Y \geq \$50{,}000 & \end{cases}
$$

where $Y50^* = (50000.^{\theta} - 1)/\theta$. The covariance parameters σ_u and ρ_{eu} are estimated directly.[22] We may then compute

$$\begin{aligned} \sigma_{eu} &= \rho_{eu}\,\sigma_u \\ \sigma_{vu} &= \sigma_{eu} - \gamma\sigma_u^2 \\ \sigma_{vv} &= 1 - 2\gamma\sigma_{vu} - \gamma^2\sigma_u^2 \\ \rho_{vu} &= \sigma_{vu}/(\sigma_u\sigma_{vv}^{1/2}). \end{aligned} \tag{2.15}$$

The model is estimated conditional on θ for values of θ which are of particular interest and in the neighborhood of the maximum. That is, we compute

$$\text{Max} \quad (\alpha,\beta,\gamma,\sigma_u,\rho_{eu},\ |X_i,Y_i,I_i|\ \theta)$$

Especially interesting values of θ correspond to the most common form of earnings functions--the log-linear earnings function ℓnY and the dollar value earnings function Y. Once the neighborhood of θ is found by estimating the other parameters conditional on θ, the parameter θ may be freed to estimate it more precisely or to estimate its standard error.[23]

III. EMPIRICAL ESTIMATES

This section summarizes a variety of empirical experiments based on the statistical approach developed in Sec. II. In this summary, we do not assume the role of advocates of any particular statistical model. Rather, our purpose is simply to explore and illustrate the sensitivity of several statistical approaches incorporating nonrandom reporting.

Our samples include employed civilian white males between the ages of 16 and 65 in each of four CPS years. We also exclude men with zero wage and salary earnings and those who were self-employed or working without pay.[24]

Given the size of the data sets used, we opt for simplicity in specifying the list of regressors in each of the behavioral relations. Earnings are assumed to be a function of education, years of market experience, the probability that a worker is in his first year of market experience, and a dummy variable for residence in the Southern states. The propensity to report earnings depends upon the level of earnings, the first-year-working probability, and the nature of the worker's demographic position in the family as indexed by two dummy variables. The first receives the value one if the worker is a child of the household head, and the second is set to one if he is some other relative of the head or a secondary family member.[25] The variables measuring type of family membership are meant to capture the possibility that a person other than the worker was interviewed. Since the head or his wife is most likely to be questioned, they may not know other family members' earnings and refuse to report. In addition, we allow the propensity to report to vary with the sample rotation month.

Because we are experimenting with alternative functional forms for earnings, we also allowed flexibility in the functional form of regressors. In the earnings equation, schooling is indexed by a set of five dummy variables: 8, 9-11, 12, 13-15, and 16+ years of schooling.[26] Similarly, years of market experience is captured by a four-segment experience spline where linear slopes are allowed to differ at breakpoints of 5, 10, and 20 years of market work. Dummy variables are used for Southern residence, child of head, other relative of head, personal interview (rotation months 1 and 5), and second-year interview wave (rotation units 5-8). Prob 1 is our estimated probability that an individual was in the first year of market work.[27] Means of all variables are given in App. Table 23 along with more precise variable definitions.

Section III is organized as follows. We first discuss our estimates of the reduced-form reporting equation. Next, we present simple OLS (not corrected for censoring), instrumental variable, and maximum likelihood estimates of the two most standard earnings functions--linear and log-linear. These two functional forms carry with them substantially different implications regarding the importance and even direction of effects for nonrandom reporting of earnings. Moreover, when we estimate the appropriate power transformation of earnings, our data strongly reject both the linear and log-linear models. Using 1980 CPS data, we estimate the "best" power transformation. Under this transformation, the effects of nonrandom income reporting are examined in all four CPS years, 1970, 1975, 1976, and 1980.

Reduced-Form Probability of Reporting Earnings

Table 14 presents our estimates of parameters of the reduced-form probit equation for the probability of reporting. For the most part, parameter estimates support a priori notions. Both secondary household members and children are significantly less likely to have earnings reported. Our interpretation is that other members of the household besides these workers were more likely to be the interviewed respondent. Since the respondent to survey questions is less certain of other members' earnings, they refuse to report. Similarly, earnings are more likely to be reported in a personal interview, but less so during the second year of the sample rotation.[28]

If reporting propensities are negatively related to income, the earnings variables in the reduced-form probit should be opposite in sign to their effects in the earnings equations we estimate below. For the most part, this is the result we obtain, but the exceptions are informative. Attending high school or beyond and acquiring additional market experience both reduce reporting probabilities.[29] The exceptions relate to the insignificant positive effects of the 8 and 9-11 years of schooling categories and the decline in reporting propensities after 20 years of experience, even though we estimate that earnings decline after this point. These exceptions may confirm the general nonreporting notion we advanced earlier. If these general nonreporters have more marginal contacts with the labor market, they are more likely to be found among men with 0-7 years of schooling (to which schooling 8 and schooling 9-11 are being compared) and among those who are nearing the end of their labor market careers.

Table 14

1980 CPS REDUCED-FORM PROBIT
(Standard Errors in Parentheses)

Earnings Equation			Reporting Equation		
Variable	Parameter	Estimate	Variable	Parameter	Estimate
Constant			Constant	$(\alpha_1 + \gamma\beta_1)$	1.0578 (.0087)
Schooling 8	$\gamma\beta_2$	.0256 (.0515)	Prob 1	$(\alpha_2 + \gamma\beta_{11})$	.0190 (.2369)
Schooling 9-11	$\gamma\beta_3$	.0066 (.0434)	Child	α_3	-.5116 (.0286)
Schooling 12	$\gamma\beta_4$	-.0084 (.0392)	Secondary Member	α_4	-.4749 (.0366)
Schooling 13-15	$\gamma\beta_5$	-.0739 (.0422)	Personal Interview	α_5	.0601 (.0200)
Schooling 16+	$\gamma\beta_6$	-.0639 (.0419)	Yr 2	α_6	-.0506 (.0171)
Experience 0-5	$\gamma\beta_7$	.0177 (.0274)	N		32,879
			$\ln \mathcal{L}$		-13,614
Experience 5-10	$\gamma\beta_8$	-.0179 (.0093)	Calculated γ		-.3494
Experience 10-20	$\gamma\beta_9$	-.0351 (.0037)			
Experience 20+	$\gamma\beta_{10}$	-.0033 (.0017)			
Prob 1					
South	$\gamma\beta_{12}$	.0330 (.0194)			

Linear and Log-Linear Models

Virtually all empirical studies of male earnings are based on linear or log-linear models. Table 15 presents our estimates for the economists' favorite, the log-linear specification, while Table 16 presents a comparable set of equations for dollar earnings. For each specification, we list OLS estimates for those reporting earnings and for the full sample including Census-imputed values. The last two columns in each table present estimates of the instrumental variable (IV) and maximum likelihood (FIML) models advanced in Sec. II.

For either functional form, estimates of parameters of the earnings equation are very similar in all four specifications. If we accept the FIML as the benchmark, the IV estimates perform least well but even they are not far off the mark. While the IV and FIML earnings equation coefficient estimates are reasonably close, the implied covariance terms differ markedly. In log or dollar form, the IV covariance terms are not feasible since the implied correlation between e and u (ρ_{eu}) is outside the range -1 to 1. With either functional form, IV estimates also proved to be extremely unstable across CPS years (in results we do not report here).

The FIML estimates (incorrectly assuming normality in both cases) are quite different for the linear and log-linear models. The log-linear model indicates a negative effect of earnings on the propensity to report earnings and a negative correlation (-.68) between the structural residuals v and u. In contrast, the linear model indicates a positive effect of earnings on the propensity to report and a positive correlation (0.86) between v and u. Similarly, for the log-linear

Table 15

1980 CPS LOG-LINEAR MODEL, $\theta \to 0$

Variable	Parameter	Reporters OLS	Full-Sample OLS	Instrumental Variable	FIML
		Earnings Equation			
Constant	β_1	9.5013 (.0039)	9.4962 (.0036)	10.0373 (.0173)	9.6816 (.0051)
Schooling 8	β_2	.2954 (.0245)	.2839 (.0227)	.2615 (.0241)	.2661 (.0202)
Schooling 9-11	β_3	.3870 (.0206)	.3963 (.0191)	.3718 (.0203)	.3692 (.0169)
Schooling 12	β_4	.6881 (.0188)	.6882 (.0174)	.6713 (.0185)	.6516 (.0158)
Schooling 13-15	β_5	.7986 (.0201)	.7898 (.0186)	.8096 (.0198)	.7694 (.0176)
Schooling 16+	β_6	1.0519 (.0199)	1.0521 (.0184)	1.0418 (.0195)	1.0445 (.0178)
Experience 0-5	β_7	-.0225 (.0119)	-.0067 (.0113)	-.0425 (.0117)	-.0294 (.0111)
Experience 5-10	β_8	.0534 (.0038)	.0546 (.0036)	.0561 (.0037)	.0557 (.0039)
Experience 10-20	β_9	.0024 (.0016)	.0221 (.0015)	.0448 (.0017)	.0240 (.0016)
Experience 20+	β_{10}	-.0052 (.0008)	-.0043 (.0008)	-.0033 (.0008)	-.0036 (.0008)
Prob 1	β_{11}	-1.8136 (.1075)	-1.6051 (.1003)	-1.5311 (.1059)	-1.7301 (.0945)
South	β_{12}	-.0654 (.0087)	-.0618 (.0082)	-.0893 (.0086)	-.0649 (.0085)

Table 15--continued

Variable	Parameter	Instrumental Variable	FIML
	Reporting Equation		
Constant	α_1		4.1137 (2.174)
Prob 1	α_2		.1839 (.0779)
Child	α_3	-.5116	-.4449 (.0182)
Rel Sec	α_4	-.4749	-.3887 (.0241)
PINT	α_5	.0601	.0475 (.0141)
Yr 2	α_6	-.0506	-.0348 (.0122)
γ		-.0917	-.3290 (.0224)
σ_u		1.3291	.7499 (.0021)
q			-1.8224 (.0209)
ρ_{eu}		-1.4864	-.9491
σ_{uu}		1.7666	.5624
σ_{eu}		-1.9759 (.0621)	-.7117
σ_{vu}			-.5267
σ_{vv}			.5926
σ_v			.7698
ρ_{uv}			-.6842

	Reporters OLS	Full-Sample OLS	Instrumental Variable	FIML
N	27,909	32,879	27,909	32,879
$\ln \mathcal{L}$				-300,613
R^2	.2716	.2705		

Table 16

LINEAR MODEL, $\theta = 1$

Variable	Parameter	Reporters OLS	Full-Sample OLS	Instrumental Variable	FIML
		Earnings Equation			
Constant	β_1	16679.1 (55.31)	16716.3 (52.6)	21346.0 (249.7)	14031.6 (64.2)
Schooling 8	β_2	2720.0 (350.1)	2694.6 (329.9)	2427.4 (348.2)	2680.1 (386.9)
Schooling 9-11	β_3	4428.1 (294.9)	4280.9 (277.8)	4259.8 (293.0)	4151.2 (333.5)
Schooling 12	β_4	7559.1 (268.7)	7559.7 (252.5)	7412.3 (267.0)	7151.2 (306.4)
Schooling 13-15	β_5	9469.1 (287.7)	9447.3 (270.7)	9564.9 (285.9)	8711.6 (318.7)
Schooling 16+	β_6	14892.3 (284.1)	1518.0 (267.4)	14804.0 (282.2)	13018.2 (314.9)
Experience 0-5	β_7	21.41 (170.9)	97.6 (163.6)	-152.8 (170.0)	129.8 (213.7)
Experience 5-10	β_8	868.3 (54.1)	865.6 (52.5)	891.4 (53.8)	765.5 (53.8)
Experience 10-20	β_9	401.4 (23.1)	403.7 (22.0)	596.7 (25.1)	304.7 (19.6)
Experience 20+	β_{10}	-86.7 (11.9)	-69.1 (11.1)	-69.7 (11.8)	-96.3 (9.8)
Prob 1	β_{11}	-11622.6 (1537.5)	-10458.1 (1458.2)	-9162.8 (1532.9)	-11824.8 (2143.9)
South	β_{12}	-867.0 (124.9)	-781.6 (119.1)	-1075.0 (124.6)	-801.8 (114.3)

Table 16--continued

Variable	Parameter	Instrumental Variable	FIML
	Reporting Equation		
Constant	α_1		.5880 (.0286)
Prob 1	α_2		-.2214 (.0930)
Child	α_3	-.5116	.0179 (.0241)
Rel Sec	α_4	-.4749	.0797 (.0241)
PINT	α_5	.0601	.0384 (.0150)
Yr 2	α_6	-.0506	-.0303 (.0127)
γ		-.0000179	.000029 (.000002)
σ_u		13698.4	9524.7 (39.0)
q			1.5929 (.0276)
ρ_{eu}		-1.2559	.9206
σ_{uu}		1.8764×10^8	40,718,767
σ_{eu}		-17.204 (19.2)	8768.3
σ_{vu}			6177.7
σ_{vv}			.5732
σ_v			.7571
ρ_{uv}			.8569

	Reporters OLS	Full-Sample OLS	Instrumental Variable	FIML
N	27,909	32,879	27,909	32,879
$\ln\mathcal{L}$				-300563.1
R^2	.2539	.2511		

model, nonreporters are predicted to earn substantially more than reporters (about twice as much), while the linear model predicts nonreporters to earn substantially less. These widely divergent results for the two most common earnings functions supply sufficient warning about the sensitivity, to inappropriate distributional assumptions, of the by now conventional sample-censoring statistical techniques.

The "Best" Power Transformation

The contradictory results produced by the standard log-linear and linear earnings equations attest to the importance of normality assumptions. Departure from normality and especially symmetry in the density of u in the assumed functional form "looks like" the effects of selectivity in the likelihood procedure conditioned on an erroneous value of θ. The dollar earnings density is known to be positively skewed while the log earnings density is slightly negatively skewed. We speculate that the FIML selectivity correction attempts to fill out each distribution to make it more normal. For log earnings, this implies adding to the upper tail so that nonreporters are given high incomes. For dollar earnings, nonreporters are assigned low incomes, filling out the lower tail.

This sensitivity led us to search for a functional form for earnings that is "best", i.e., most normal, within a wider class of functions. We have estimated the appropriate transformation in a manner similar to the Box-Cox procedure. Initial parameter estimates are obtained using the instrumental variable procedure for a grid of values of of θ.[30] The maximum likelihood procedure for the full model,

including the reporting equation, is used to estimate parameters conditional on θ; that is,

$$\text{Maximize } \ell n\, L(\alpha,\beta,\gamma,\sigma_u,\rho_{eu} \mid X,I,Y;\ \theta).$$

A subsample of observations in the first rotation group, all personal interviews, was used for the first round of estimates exploring a substantial number of θ values. The results of this exploration are reported in Table 17. The conditional likelihood function appears to be concave in θ (see Fig. 2) with a maximum at $\theta = 0.45$ (considering intervals of 0.05). The same maximum appears to be valid when we expand our estimation to the full sample. Both the linear and log-linear models are strongly rejected.[31]

Parameter estimates for $\theta = 0.45$ are presented in Table 18 for our four CPS years.[32] In this functional form, earnings equation estimates and parameter estimates from the equation for reporting propensity are now reasonably stable over the years. Although the estimate of the effect of earnings on the probability of reporting, γ, varies little from year to year, there is a growing negative correlation between earnings and the propensity to report earnings, through the correlation in residuals ρ_{vu}. These negative correlations imply that in our best normal distribution, nonreporters of earnings have larger incomes than similar reporters.

Table 17

LOG LIKELIHOOD CONDITIONAL ON θ

θ	Rotation 1 N = 4080	Full Sample N = 32,879
.00	-38,206.66	-300,613.40
.25	-37,893.24	
.25	-37,893.24	
.30	-37,864.71	
.35	-37,845.04	
.40	-37,833.43	-298,180.29
.45	-37,829.25	-298,169.68
.50	-37,831.97	-298,207.95
.75	-37,951.86	
1.00	-38,140.05	-300,563.10

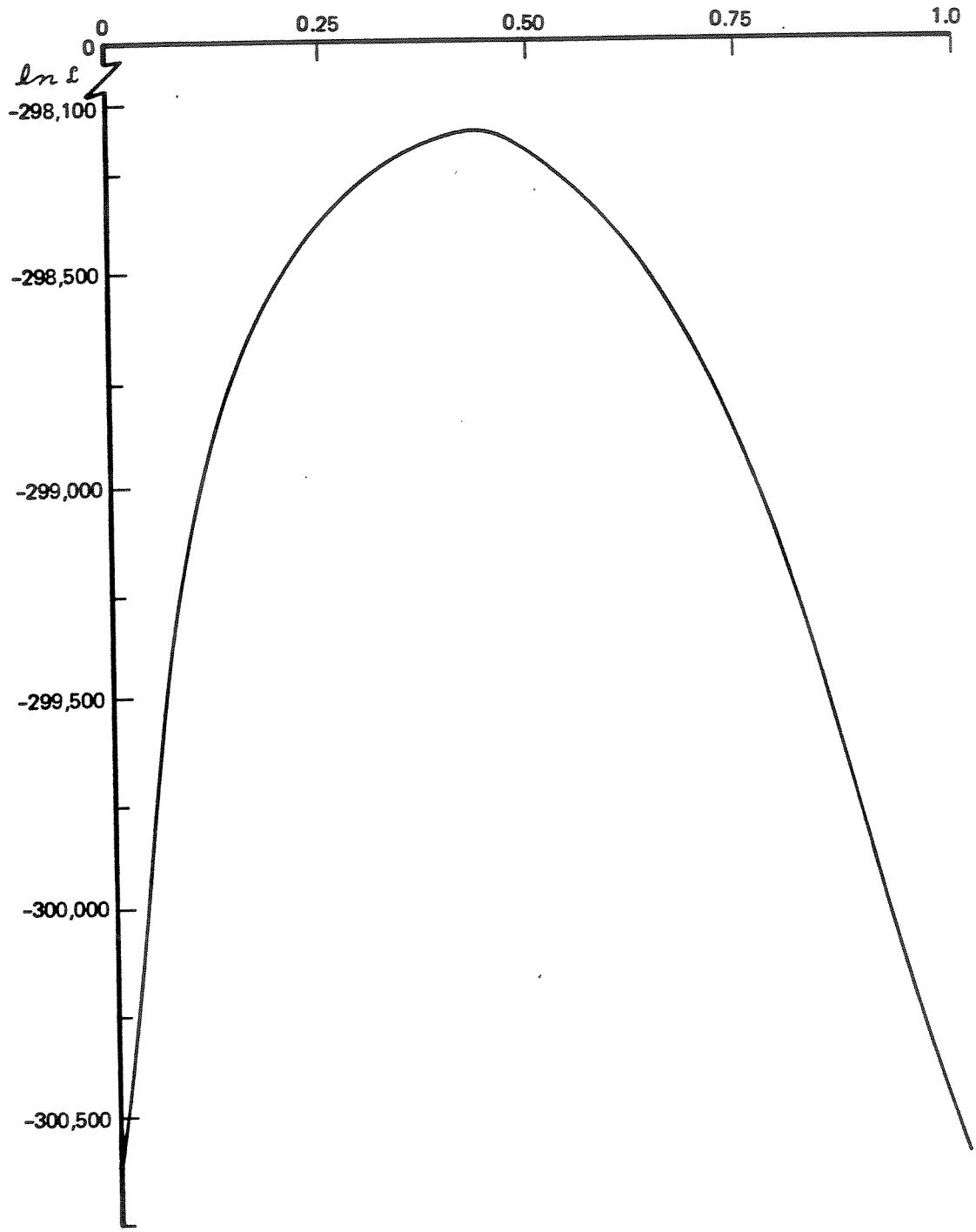


Fig. 2 — Graph of $\ln \mathcal{L}\ (\alpha, \beta, \gamma, \sigma_u, \rho_{\epsilon u} \mid X, I, Y; \theta)$

Table 18

MAXIMUM LIKELIHOOD ESTIMATES CONDITIONAL ON $\theta = .45$

Variable	Parameter	1970	1975	1976	1980
		Earnings Equation			
Constant	β_1	173.8631 (.5355)	177.3678 (.7522)	173.4840 (.4279)	173.6730 (.3555)
Schooling 8	β_2	14.0397 (1.1855)	13.9620 (1.6551)	17.3731 (1.7015)	16.9415 (1.5112)
Schooling 9-11	β_3	23.6291 (1.0644)	24.1687 (1.4539)	24.5565 (1.3897)	24.7704 (1.2807)
Schooling 12	β_4	37.1995 (.9791)	40.3074 (1.3452)	42.6343 (1.2711)	42.6064 (1.1730)
Schooling 13-15	β_5	49.4214 (1.0935)	52.7007 (1.4564)	55.0063 (1.3899)	51.0225 (1.2519)
Schooling 16+	β_6	75.0080 (1.0682)	76.8700 (1.4375)	80.1619 (1.3685)	71.9481 (1.2296)
Experience 0-5	β_7	1.0774 (1.0544)	1.9038 (1.0556)	.8580 (1.0547)	-.4304 (.8352)
Experience 5-10	β_8	3.0021 (.2867)	4.5811 (.3078)	4.5217 (.3015)	4.0063 (.2463)
Experience 10-20	β_9	1.0755 (.1019)	1.4864 (.1165)	1.6703 (.1156)	1.7214 (.0980)
Experience 20+	β_{10}	-.3894 (.0419)	-.4046 (.0524)	-.3722 (.0559)	-.3408 (.0483)
Prob 1	β_{11}	-110.1410 (9.9960)	-78.2636 (9.8567)	-82.2772 (9.7561)	-85.9818 (7.8190)
South	β_{12}	-8.2455 (.5372)	-5.0754 (.6095)	-4.6133 (.6192)	-4.6617 (.5408)

Table 18--continued

Variable	Parameter	1970	1975	1976	1980
		Reporting Equation			
Constant	α_1	1.9368 (.0860)	1.8261 (.0778)	1.6512 (.0696)	1.5075 (.0677)
Prob 1	α_2	.5773 (.1601)	.4753 (.1318)	.4868 (.1167)	.2382 (.0977)
Child	α_3	-.5909 (.0467)	-.5689 (.0436)	-.6620 (.0384)	-.5874 (.0265)
Rel Sec	α_4	-.1038 (.0745)	-.0472 (.0650)	-.5275 (.0454)	-.5490 (.0348)
PINT	α_5	.1600 (.0257)	.0830 (.0230)	.0794 (.0208)	.0605 (.0180)
Yr 2	α_6	-.1147 (.0211)	-.0255 (.0194)	-.0138 (.0176)	-.0441 (.0154)
γ		-.0039 (.0005)	-.0045 (.0004)	-.0041 (.0004)	-.0027 (.0004)
σ_u		35.8343 (.2636)	39.8103 (.3700)	41.9923 (.2843)	43.2598 (.2438)
q		-.2074 (.0569)	-.4886 (.0596)	-.8564 (.0260)	-.9036 (.0258)
ρ_{eu}		-.2044	-.4531	-.6944	-.7181
σ_{uu}		1284.0958	1584.8607	1763.3541	1871.4086
σ_{eu}		-7.3260	-18.0372	-29.1591	-31.0627
σ_{vu}		-2.3369	-10.8882	-21.9878	-25.9181
σ_{vv}		.9625	.8695	.7920	.8434
σ_v		.9810	.9325	.8899	.9183
ρ_{vu}		-.0665	-.2933	-.5884	-.6524
N		25,048	23,707	23,947	32,879
$\ln \mathcal{L}$		-238,537.9	-216,388.9	-213,539.6	-298,169.7

Implications of the Estimates for the Distribution of Earnings

In this section, we report a number of simulations based on our preferred estimates for 1980. Predicted values of the propensity to report earnings, Z(x), and the probability of not reporting, Pr(NR) = $\Phi(-Z(X))$, are listed in the second and third columns of Table 19. The mean probability approximates the actual proportion not reporting to within one-half percentage point. However, the estimated probabilities of not reporting for actual reporters and actual nonreporters differ by only 1.54 percentage points. While the relationship is statistically significant, the difference is not very impressive. While the difference in mean probabilities is small, it does have a large impact on mean earnings. Earnings directly affects reporting, and the two are strongly negatively related because of unmeasured variables represented by σ_{uv}.

We next compute values of the transformed values (i.e., $Y^*(X) = \beta'X$). Strictly on the basis of the measured variable X, nonreporters have a slightly higher predicted value of $Y^*(X)$, but the expected values controlling for reporting status are substantially different. Since dollar values are easier to understand, we list expected dollar values as well. Dollar values are a non-linear function of Y* given by

$$Y = (\theta Y^* + 1)^{1/\theta} = (0.45(Y^*(X)+u)+1)^{1/0.45} \quad .$$

First, values of $Y(X) = (0.45Y^*(X)+1)^{1/0.45}$ are computed for each individual and averaged. This is not an estimate of the mean, but

Table 19

SELECTED MEAN COMPUTATIONS FROM 1980 PARAMETER ESTIMATES (θ = .45)

	Reporting		Transformed		Dollars			
Sample	$\hat{Z}(X)$	R-(NR)	$\hat{Y}^*(X)$	$E(\hat{Y}^*(X \mid I))$	$\hat{Y}(X)$	$E(Y(X))$	$E(Y(X \mid I))$	CPS $
All	1.030	.1561	173.67	173.42	17,060	18,415	18,375	16,717
Reporters	1.039	.1537	173.49	164.83	17,060	18,365	16,343	16,634
Nonreporters	.981	.1691	174.70	221.69	17,337	18,695	29,787	17,183

rather the average of the function of individual mean values, and is reported only as an intermediate step.

Expected values of Y conditional on X may be computed directly from the moments of u by writing the function as a binomial series,[33] recognizing that odd moments of the centralized normal are zero, i.e.,

$$E(Y|X) = (.45\hat{Y}^*(X)+1)^{1/.45} \sum_{i=0}^{\infty} \binom{1/.45}{2i} \left(\frac{(.45)^2 \sigma_u^2}{(.45\hat{Y}^*(X)+1)^2} \right)^i .$$

This may be interpreted as applying a variance correction factor to $Y(X)$.[34] The correction will be positive for $\theta = 0.45$.[35] For the full sample, the estimate of the mean value of Y is \$18,415, which is a substantial increase over the reported mean earnings in the CPS of \$16,717. Thus, accounting for nonrandom income, nonreporting implies a 10.02 percent increase in mean earnings for all 1980 white male wage-earners.

That is a substantial increase. How can we verify to some extent whether the model is working? CPS earnings values include imputed earnings for nonreporters and actual values for reporters (using \$75,000 for the over \$50,000 category). Since earnings are observed for reporters, we may compare them directly with estimated values from the model. This requires computation of the expected value of dollar earnings for each worker, conditional on his reporting his earnings status.

The expected value of earnings Y conditional on reporting earnings (I=1) for those who reported earnings may again be computed by using the binomial series and taking expectation so that

$$E(Y|X,I=1) = (.45\hat{Y}^*(X)+1)^{1/.45} \sum_{i=0}^{\infty} \binom{1/.45}{i} \left(\frac{.45\,\sigma_u}{(.45\hat{Y}^*(X)+1)}\right)^i E\left(\left(\frac{u}{\sigma_u}\right)^i \,\middle|\, \varepsilon > -\hat{Z}(X)\right).$$

The moments of u/σ_u given $e > -Z(X)$ are related in a simple recursive way which is easy to compute. This expectation was computed for reporters and appears in Table 19. The mean of these expected values for reporters is \$16,343, which is within 2 percent of the actual mean of \$16,634. Given the complexity of the computation and its possible sensitivity to the normality assumption, the model is predicting remarkably well for reporters. For nonreporters, we estimate that the Census imputation method understates their income by 73 percent in 1980.

Similar simulations are computed for the other CPS years in Table 20. The results for all years are basically similar to those obtained for 1980. The difference in estimated mean earnings between this procedure and the CPS value is 11 percent in 1976, 7.6 percent in 1975, and only 1 percent in 1970. Using the model, real earnings fell 3.6 percent between 1975 and 1976. Using the CPS mean values, real earnings fell 6.5 percent. Our simulations also predict that the bias in the Census imputation method has grown over time. Comparing our estimate for nonreporters with Census-assigned values, the Census understates income of nonreporters by 15.0 percent (1970), 44.0 percent (1975), 63.0

Table 20

SELECTED MEAN COMPUTATIONS FROM 1975 and 1976 PARAMETER ESTIMATES ($\theta = .45$)

		Reporting		Transformed		Dollars			
Sample	N	$\hat{Z}(X)$	R-(NR)	$\hat{Y}^*(X)$	$E(\hat{Y}^*(X \mid I))$	$\hat{Y}(X)=$ $(OY^*(X)+1)^{1/\theta}$	$E(Y(X))=$ $E(OY^*(X)+1)^{1/\theta}$	$E(Y(X \mid I))$	CPS $
1976									
All	23,947	.946	.176	173.48	173.28	17,114	18,391	18,361	16,550
Reporters	19,823	.956	.174	172.70	163.68	16,942	18,218	16,121	16,276
Nonreporters	4,124	.895	.190	177.25	219.38	17,941	19,223	29,125	17,868
1975									
All	23,707	1.026	.156	177.37	177.29	17,925	19,078	19,063	17,710
Reporters	20,075	1.033	.154	176.55	171.51	17,743	18,895	17,717	17,589
Nonreporters	3,632	.986	.165	181.89	209.23	18,931	20,090	26,504	18,384
1970									
All	25,048	1.261	.107	173.86	173.85	17,059	17,990	17,988	17,876
Reporters	22,395	1.267	.106	173.45	171.96	16,969	17,899	17,570	17,776
Nonreporters	2,653	1.213	.117	177.33	189.78	17,823	18,758	21,519	18,715

percent (1976), and 73.4 percent (1980). This pattern is consistent with the public's alleged growing sensitivity to marginal tax rates during the 1970s.

In Table 21, we consider the distribution of earnings in greater detail to further assess the fit of the model and to gain further inferences from the results. The table presents the relative frequency distributions of CPS and estimated earnings for 1980. Columns (1)-(3) report the frequency distributions using CPS values. Columns (4)-(6) report the corresponding estimated values. Column (7) reports the estimated distribution using actual values for reporters and estimated values for nonreporters.

The most relevant comparison is for reporters, since CPS values are actual values. As noted earlier, the means differ little. Column (8) reports the difference in proportions. The largest difference is just under 3 percent, but we are clearly not capturing the real distribution exactly. This may reflect a sharp sensitivity to θ, and the fact that we only explore values that are multiples of 0.05. The value of $\theta = 0.45$ may not yet be accurate enough. A change in $(-\ell n)$ of 2.0 would be significant at the 5 percent level and the nearest value $\theta = 0.40$ yields a change of 10.4. This should be explored in further work.[36]

Does the Change in Census Imputation Methods Matter?

Table 1 listed the significant changes in the Census imputation method that took place in 1976. Since many researchers have used CPS data to track changes over time, this raises the question of whether Census alterations in handling nonreporters affects interpretation of

Table 21

CPS AND ESTIMATED EARNINGS DISTRIBUTIONS, 1980

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Percent Frequency							
	Sample			Estimated				
Income Interval $1,000	Full Sample	Reporters	Non-Reporters	Full Sample	Reporters	Non-Reporters	Actual Reporters + Simulated Nonreporters	Reporters (Estimated - Actual)
0-3	4.30	4.09	5.43	3.38	3.35	3.52	4.01	-.74
3-6	7.01	6.81	8.17	6.96	6.97	6.90	6.82	.16
6-9	9.92	9.85	10.26	9.84	9.89	9.58	9.81	.04
9-12	12.79	12.91	12.09	11.48	11.55	11.14	12.64	-1.36
12-16	18.49	18.79	16.78	15.74	15.81	15.34	18.67	-2.98
16-19	12.81	13.11	11.17	10.92	10.95	10.75	12.75	-2.16
19-25	18.59	18.68	18.09	17.52	17.52	17.52	18.50	-1.16
25-30	7.69	7.65	7.91	9.84	9.80	10.05	8.00	2.15
30-35	3.53	3.58	3.28	6.26	6.22	6.52	4.02	2.64
35-40	1.87	1.82	2.13	3.72	3.68	3.94	2.14	1.86
40-50	1.66	1.52	2.46	3.21	3.16	3.47	1.81	1.64
50+	1.35	1.20	2.23	1.12	1.10	1.26	1.20	-.10
N	32,879	27,909	4,970	32,879		4,920	32,879	

secular trends. To make the application more concrete, we concentrate here on the estimated income returns from schooling.

Table 22 lists alternative earnings functions for the 1975 and 1976 CPS samples. We estimate three OLS earnings equations for each year: for reporters and nonreporters separately, and for the combined sample of reporters and nonreporters. For comparison, we also list our FIML equations in both years. Between these two years, the Census radically altered their imputation algorithm. In particular, schooling was included as a control to find matches in 1976, but it was not in 1975.

Table 22 indicates that such alterations in the Census methodology do, in fact, matter. As one would expect, the estimated income benefits from schooling for 1975 nonreporters of income are substantially below those of reporters. Since schooling was not used to obtain their incomes, this pattern for 1975 nonreporters is a statistical artifact. Because the OLS on the full samples averages the equations for reporters and nonreporters, we would more accurately approximate the 1975 FIML with an OLS over reporters.

In 1976, however, the differences in the OLS estimates for reporters and nonreporters are considerably smaller. Therefore, researchers attempting to track changes over time with CPS data are best off confining their analysis to income reporters.

Table 22

COMPARISON OF EARNINGS FUNCTIONS (FOR $\theta = .45$) AT BREAKPOINT IN CHANGES IN CENSUS ALGORITHM

	1975 CPS				1976 CPS			
Variables	OLS Full Sample	OLS Reporters	OLS Nonreporters	FIML	OLS Full Sample	OLS Reporters	OLS Nonreporters	FIML
Constant	170.6 (.267)	171.1 (.282)	167.6 (.832)	177.4 (.752)	165.1 (.260)	164.9 (.278)	166.3 (.727)	173.5 (.430)
Schooling 8	12.8 (1.54)	13.7 (1.61)	6.67 (4.58)	13.9 (1.66)	15.9 (1.49)	16.6 (1.58)	10.1 (4.17)	17.4 (1.70
Schooling 9-11	22.4 (1.34)	23.4 (1.41)	1.49 (3.97)	24.2 (1.45)	22.2 (1.25)	23.3 (1.33)	14.8 (3.60)	24.6 (1.40)
Schooling 12	37.6 (1.24)	39.5 (1.30)	25.3 (3.71)	40.3 (1.35)	40.0 (1.15)	40.8 (1.21)	33.9 (3.30)	42.6 (1.27)
Schooling 13-15	49.2 (1.36)	51.7 (1.43)	34.5 (4.01)	52.7 (1.46)	51.7 (1.27)	52.8 (1.34)	44.0 (3.59)	55.0 (1.39)
Schooling 16+	71.6 (1.34)	76.2 (1.41)	45.7 (3.99)	76.9 (1.44)	76.9 (1.24)	77.1 (1.32)	73.5 (3.49)	80.2 (1.37)
Experience 0-5	2.94 (.923)	1.71 (.945)	13.4 (.322)	1.90 (1.06)	1.02 (.873)	9.88 (.915)	1.32 (2.62)	.858 (1.05)
Experience 5-10	4.16 (.289)	4.51 (.294)	1.98 (1.05)	4.58 (.308)	4.49 (.270)	4.38 (.282)	5.19 (.827)	4.52 (.302)
Experience 10-20	1.41 (.114)	1.43 (.119)	1.80 (.356)	1.49 (.117)	1.50 (.110)	1.56 (.117)	1.13 (.315)	1.67 (.116)
Experience 20+	-.364 (.054)	-.458 (.058)	-.087 (.143)	-.406 (.052)	-.348 (6.51)	-.398 (.058)	-.133 (.135)	.372 (.056)
Prob 1	-73.9 (8.20)	-80.5 (8.47)	-7.96 (27.0)	-78.3 (9.86)	-80.8 (7.80)	-82.1 (8.22)	-72.6 (3.16)	-82.3 (9.76)
South	-3.91 (.596)	-4.29 (.620)	-1.40 (1.87)	-5.08 (.609)	-3.90 (6.76)	-4.15 (.611)	-2.45 (1.62)	-.461 (.619)
N	23707	20075	3632	23707	23,947	19,823	4,124	23,947
R^2	.308	.331	.229		.278	.337	.290	

FOOTNOTES

1. This research was supported by a grant from the National Science Foundation. We gratefully acknowledge the cooperation and excellent advice of Chuck Nelson and John Coder, both of the Census Bureau. We also are indebted to Karl Schutz and Christina Witsberger of Rand for expert programming assistance.

2. Lest anyone feel we are casting stones unjustly, at least two of us admit to being among the greatest sinners of all.

3. This is not strictly true. At an earlier stage, some allocations are made for longest job (equating it with current job, if known) and work experience. If an allocation for these two variables was not made at these earlier stages, the three variables are then allocated jointly.

4. Nonlabor income is imputed separately.

5. Current members of the armed forces, and those whose main reason for not working last year was military service, are allocated separately using an algorithm that does not correspond to the one described in the text.

6. If non-response is related to income, as we argue here, this secular trend is not surprising.

7. As we shall see below, Table 3 will understate the true difference in incomes if the probability of responding is itself a function of income.

8. Published Census data indicate even higher fractions of non-reporting among the self-employed. However, we have found an error in the public-use tapes for people with self-employment income. Because of a mistake in their Census computer program, all individuals with negative incomes are flagged as having allocated incomes. If we included men with negative incomes as non-reporters, the proportions in Table 5 would change to 30.9 for self-employed non-farm and 35.7 for self-employed farm income. Unfortunately, this misclassification has been part of the Census procedure from its inception. In personal correspondence, we understand that this will be corrected starting with the 1982 CPS.

9. However, Table 5 indicates no effect regarding personal vs. telephone surveys during the final four months.

10. One pattern in Table 5 that puzzles us is the decline in reporting proportions in the final month of each stage (months 4 and 8).

11. While this variance-preserving property is desirable, if the propensity to report is nonrandom, observations on reporters will yield biased estimates of true residual variances.

12. Once again, we are assuming for the present that refusals are random.

13. We omit education and age from this table because they are always used to match. However, even for these two variables, the majority of matches are not made at the full disaggregation listed in Table 1.

14. To some extent, the marital status and labor force status of wife variables are substitutes, as the latter contains an implicit marital status control. However, the labor force variable does not distinguish between the categories that make up the non-married group.

15. Obviously, we have selected two occupations that are most likely to exhibit the most dramatic differences. This is partly because the list of 46 occupation groups lumps lawyers and judges with a hodge-podge of other professionals. Nevertheless, Table 12 effectively illustrates the consequences of relying on Census matching rather than a variable's contribution to explaining income differences.

16. The only mean difference worth noting in Table 13 is that nonreporters earning less than $3000 dollars work about one and a quarter weeks more than reporters. This may indicate that the Census is understating income for those nonreporters.

17. The reader will note that the fraction of exact matches at either 3-digit occupation code, Census region, or North-South is considerably

higher in Table 13 than in Table 11. Table 11 listed the proportion of cases in which the Census explicitly used a characteristic. Even if the Census does not use a characteristic, some individuals will match with their donors on a purely random basis. For example, if 25 percent of whites lived in the South, a completely random matching of white male nonreporters with other males on the CPS would match 62.5 percent of them on a North-South division. Therefore, the informational content of the Census procedure, which matches 85 percent on this criteria, should be measured relative to this random assignment base.

18. A statistical model similar in spirit although not in its detail was independently developed in Greenlees, Reece, and Zieschang (1982).

19. If there were no issue of selectivity, this would be a standard application of the procedure developed by Box and Cox (1964). For an application to earnings functions, see Heckman and Polachek (1974). Hernandez and Johnson (1980) show that the maximum likelihood parameter estimates u coverage to estimates that minimize the Kullback-Leibler information number with respect to the normal distribution. That is, they converge to the "most normal" distribution by the Kullback-Leibler information number criterion.

20. For $\theta > 0$, values of Y^* are limited by $Y^* > -\theta^{-1}$. We ignore this truncation since empirically the mean will be four standard deviations from the limit.

21. Note that the implicit normalization has already been made that the

cutoff for reporting is zero. This cutoff and the intercept of the Z equation are not separately identified.

22. Actually, q is estimated such that $\rho_{eu} = \tanh(q)$ to maintain $-1 < \rho_{eu} < 1$ during search. The Berndt, Hall, Hall and Hauseman (1974) algorithm is used to search for the maximum.

23. In practice, the parameters β, γ and σ_u are greatly affected by changes in θ so that a grid search is much more computationally efficient and guards as well against local optimum. In the same view, search algorithms often attempt to take steps leading to parameter values violating the Cauchy-Schwartz Inequality. Using $\rho_{eu} = \tanh(q)$ solves this problem.

24. In addition, we eliminated men (1) whose major activity last week was school, (2) who worked part-year if the reason was school, retirement, or the armed forces, or (3) who had a computed weekly wage of less than $10.

25. Other types of family membership positions were explored, but found to be insignificantly different from the household head, the omitted category.

26. The left-out group is 0-7 years of schooling.

27. For a derivation of Prob 1, see Smith and Welch (1978).

28. Our proxies for these two effects are not exact. Replacement households are given the rotation month of the household they replaced. In addition, replacement households are always given a personal interview during their initial month.

29. The inclusion of the Prob 1 variable complicates interpretation of the experience profile of earnings. In the earnings functions we report below, Prob 1 has a powerful negative effect on earnings. However, incremental years of work in the first five-year segment of the experience spline actually reduce earnings. Since the probability of being in the first year of market work declines rapidly over the first five years of experience, the effects of early years of market work are best read from the joint effects of the Prob 1 and exp 0-5 variables. If we simulate their effects together, earnings rise most rapidly over the first five years of work and the probability of reporting earnings declines. After digesting our results, we must confess that if we had to start from scratch, we would make the world simpler by dropping the Prob 1 variable from the analysis.

30. Since only one interval is available for earnings values over $50,000, a value of $75,000 is assigned to observations in that interval. The likelihood procedure fully accounts for the interval observation.

31. A 95 percent confidence interval on θ (using $-2\ell n \sim \chi^2$) does not include the values 0.40 and 0.50. The value of θ could be estimated

more precisely by estimating the unconditional likelihood. There is some evidence that the conditional likelihood approach may overstate the precision of other parameter estimates--see Bickel and Doksum (1981)--but here there are enough observations so that precision is not a problem. It is of some interest to note that the log likelihoods are quite similar for the log and linear specifications. This is in stark contrast to results reported by Heckman and Polachek (1974). In their work, the log specification was strongly preferred to the linear. We suspect that our flexibility in the functional form of regressors provides a fairer test between these models.

32. Earnings in each year are in 1979 dollars.

33. The series converges as long as $0.45\ \sigma_u < (0.45Y^*(X)+1)$. Empirically, terms become smaller than 0.0001 by the eighth term in the series.

34. As far as we know, this is an original result. The resulting correction factor is different from the expansion suggested by other authors, such as Neyman and Scott (1960). A similar formula arises for arbitrary moments.

35. The series is finite for values of θ equal to the reciprocal of integers. This transformation is close to the square of Y^* so that corrections are similar to applying variance to a parabolic function argument. The correction for the log normal distribution occurs when $\theta \rightarrow 0$.

36. Alternatively, the family of power transformations may prove inadequate.

REFERENCES

Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hauseman. 1974. Estimation and inference in nonlinear structural models. Annals of Economic and Social Measurement, 3, 4: 653-665.

Bickel, P. J., and K. A. Doksum. June 1981. An analysis of transformation revisited. Journal of the American Statistical Association, 76, 374: 296-311.

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society, B26, 2: 211-252.

Greenlees, J. S., W. S. Reece, and K. D. Zieschang. 1982. Imputation of missing values when the probability of response depends upon the variable being imputed. Journal of the American Statistical Association.

Heckman, J. J. 1980. Sample bias as a specification error. In James Smith, ed., Female Labor Supply: Theory and Estimation, Princeton University Press.

Heckman, J. J., and S. Polachek. June 1974. Empirical evidence on the function form of the earnings-schooling relationship. Journal of the American Statistical Association, 69, 346: 350-354.

Hernandez, F., and R. A. Johnson. December 1980. The large-sample behavior of transformations to normality. Journal of the American Statistical Association, 75, 372: 855-861.

Neyman, J., and E. L. Scott. 1960. Correction for bias introduced by a transformation of variables. Annals of Mathematical Statistics, 31: 643-655.

Poirier, D. J., and A. Melino. September 1978. A note on the interpretation of regression coefficients within a class of truncated distributions. Econometrica, 46, 5: 1207-1209.

Poirier, D. J., June 1978. The use of Box-Cox transformation in limited dependent variable models. Journal of the American Statistical Association, 73, 362: 284-287.

Smith, J. P., and F. R. Welch. 1978. Cyclic components in the demand for college trained manpower. Annals D'INSEE, Vol. 30-31.

A P P E N D I X

Table 23

SUMMARY OF INCOME IMPUTATION HISTORY

CPS Years	Treatment of Nonresponses
1948-1961	Not included in published data.
1962-1965	Imputed all income categories if one income item was nonresponse. Variables used for imputation were age, weeks worked, race, and major occupation group.
1966-1967	Imputed only for missing income item--same variables as for 1962-1965.
1968-1975	Added sex and type of family member to impute.
1976-1978	Added education, labor force status of spouse, number of children, marital status, region, urbanization, class of worker, and full-time/part-time status.

NOTE: In addition to expanding the number of variables used in income imputation, the CPS periodically added more detail to the variables used.

Table 24

VARIABLE MEANS AND DEFINITIONS

Variable	Means by Year 1970	1975	1976	1980	Definition
I	.894	.847	.828	.849	= 1 if report earnings; 0 otherwise
Schooling 0-7	.081	.057	.064	.051	= 1 if years of school is 7 or less; 0 otherwise
Schooling 8	.099	.066	.059	.051	= 1 if years of school is 8; 0 otherwise
Schooling 9-11	.172	.144	.143	.127	= 1 if years of school is 9-11; 0 otherwise
Schooling 12	.369	.389	.388	.389	= 1 if years of school is 12; 0 otherwise
Schooling 13-15	.130	.157	.158	.177	= 1 if years of school is 13-15; 0 otherwise
Schooling 16+	.148	.187	.188	.205	= 1 if years of school is 16 or more; 0 otherwise
Experience 0-5	4.678	4.498	4.488	4.466	= years of experience if less than 5; 5 otherwise
Experience 5-10	3.953	3.635	3.584	3.468	= years of experience beyond 5 if between 5 and 10; 0 if less than 5; 5 if more than 5
Experience 10-20	6.042	5.395	5.226	4.794	= years of experience beyond 10 if between 10 and 20; 0 if less than 10; 10 if more than 20
Experience 20+	5.328	4.792	4.541	4.004	= years of experience beyond 20 if beyond 20; 0 if less than 20
Prob 1	.027	.041	.042	.041	
South	.269	.287	.292	.269	= 1 if region is South; 0 otherwise
Child	.070	.087	.091	.120	= 1 if child of household head; 0 otherwise
Rel Sec	.017	.019	.018	.020	= 1 if relative of head or secondary household member; 0 otherwise
PINT	.244	.248	.250	.249	= 1 if rotation month 1 or 5 which implied a personal interview; 0 otherwise
Yr 2	.498	.497	.497	.497	= 1 if rotation month 5 to 8 which is second year; 0 otherwise