Evaluating a Multiple-Imputation Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard

Donald J. Treiman; William T. Bielby; Man-Tsun Cheng

Sociological Methodology, Vol. 18. (1988), pp. 309-345.

Stable URL:
http://links.jstor.org/sici?sici=0081-1750%281988%2918%3C309%3AEAMMFR%3E2.0.CO%3B2-C

Sociological Methodology is currently published by American Sociological Association.

10

Evaluating a Multiple-Imputation Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard

Donald J. Treiman, William T. Bielby† and Man-tsun Cheng‡*

In this paper, we evaluate a multiple-imputation procedure for converting data coded into the 1970 *U.S. census detailed classification of industries to the categories of the* 1980 *classification. For a sample of* 127,125 *persons in the labor force drawn from responses to the U.S. census of* 1970, *we compare* 1980 *codes directly assigned by coders working from the narrative descriptions in the census questionnaires with* 1980 *codes imputed via a logistic regression procedure from the* 1970 *codes and other information about*

Preparation of this paper was supported by National Science Foundation grant SES 83-11483. An earlier version of the paper was presented at the 1987 meetings of the American Sociological Association, Chicago, and the final version was prepared while Treiman was an ASA/NSF/Census Fellow. Clifford Clogg, Robert Johnson, Ann Miller, Charles Nam, John Priebe, Patricia Roos, Donald Rubin, Nathaniel Schenker, Tom Scopp, Paul Siegel, Lynn Weidman, and three anonymous referees provided helpful comments on an earlier version. We are grateful to the staff of the UCLA Social Sciences Computing Facility for support in data processing.

* University of California at Los Angeles.

† University of California at Santa Barbara.

‡ University of California at Los Angeles.

respondents. By replicating the imputation several times—in this case, five —and appropriately combining the multiple imputations, we can estimate the additional uncertainty introduced by the imputation procedure. We show that the additional error due to imputation tends to be small relative to sampling error and that using imputed data in large (2 *percent*) *Public-Use Microdata Samples is preferable to using directly assigned data from the smaller* 127,125 *case sample. We offer various cautions regarding the appropriate use of imputed data.*

1. INTRODUCTION

Students of social change are perennially plagued by problems of data noncomparability. The very changes that are the object of study ordinarily engender changes in the way the data necessary to analyze change are collected. Industrial and occupational data are prime examples. Each decade the U.S. Census Bureau collects relatively detailed data on the industrial and occupational composition of the U.S. labor force.[1] But for each census, the detailed industry and occupation classifications are changed to reflect salient variations among industries and occupations at the time of data collection. Thus, new industries and occupations are added and declining industries and occupations are deleted. Moreover, categories are sometimes recombined in complex ways.

For example, the 1960 census contained no specific occupation category for computer programmers, including the 8,700 or so programmers in the labor force at that time in the residual category, "professional, technical, and kindred workers, n.e.c." (Priebe, Heinkel, and Greene 1972, p. 191). By 1970 there were 263,000 computer specialists, divided into three categories in the 1970 census classification: "computer programmers," "computer systems analysts," and "computer specialists, n.e.c." (U.S. Bureau of the Census 1973, p. 1). In 1980 the classification was changed still again, this time reducing the number of categories to two: "computer systems analysts and

[1] *Occupation* refers to the characteristics of a job—the function fulfilled and the tasks performed to accomplish it. Examples of occupations include carpenter, secretary, and physician. *Industry* refers to the goods or services produced by the enterprise within which the job is performed. Examples of industries are automobile manufacturing, construction, and educational services.

scientists" (a subcategory of "mathematical and computer scientists") and "computer programmers" (a subcategory of "technicians, except health, engineering, and science") (U.S. Bureau of the Census 1981, pp. xi, xiii). The 1970 category "computer programmers" was split between the two 1980 categories, but the two other 1970 categories mapped only into the 1980 category "computer systems analysts and scientists" (Vines and Priebe 1988, Table 1).

Many other recombinations were even more complex, and fewer than one third of the occupation categories in the 1970 detailed occupation classification mapped into a single category in the 1980 detailed occupation classification (Vines and Priebe 1988, Table 1). Moreover, the categories least likely to change tended to be those involving small numbers of workers, so that less than 15 percent of the labor force in 1970 was in a category that mapped into a single 1980 category (computed from Vines and Priebe 1988, Table 1).

Over the past several censuses, changes in the classification of detailed industries have not been as extensive as changes in the detailed occupation classification (because of the adoption of the *Standard Industrial Classification* prior to the 1940 census), but they are still substantial. About 24 percent of the 1970 industry categories mapped into more than one 1980 category classification, and the non-matching categories included about 36 percent of the 1970 labor force (Vines and Priebe 1988, Table 3).

While the decadal changes in the detailed classification schemes for industry and occupation have little practical consequence for cross-sectional analysis, they pose severe difficulty for many kinds of cross-temporal analysis. For example, it is currently impossible to assess adequately from census data the effect of a decade of affirmative action, since it is impossible to answer accurately such questions as, "Has the proportion of women in management changed between 1970 and 1980?" The difficulty is that there is no way of knowing whether a given job would be counted as a managerial occupation in one scheme but not in the other. Indeed, a person who had the same job in 1970 and 1980 might have moved into or out of a detailed occupational category that we would be willing to count as a managerial occupation simply because of a change in the occupational coding scheme. Similar problems plague attempts to assess changes in the industrial distribution of the labor force. For example, assessment of the claim that jobs in secondary industries are increasingly occupied by women, minorities,

and immigrants depends on consistent classification of industries over time.

The Census Bureau has been concerned with the problem of cross-temporal comparability of industry and occupation classifications for a very long time. The census of 1900, for example, includes a comparison of occupation data from 1820 to 1900, although these comparisons were somewhat limited (U.S. Bureau of the Census 1904). In the early 1940s, Alba Edwards undertook a major monographic study on occupational comparability since 1870 (Edwards 1943). In addition to data for 1940, Edwards's monograph provided detailed occupational data for 1930, reclassified into the 1940 classification, and comparable data for 1870 through 1930, classified according to the 1930 classification. This was followed by a similar effort by Kaplan and Casey (1958), which provided detailed occupational distributions for males and females for 1900 through 1950, classified according to the 1950 classification. The Edwards and Kaplan and Casey monographs each provided extremely useful data because they permitted estimates of historical change in the occupational structure that were uncontaminated by changes in the classification scheme. But, of course, they permitted no detailed analysis of change in the occupational composition of the labor force with respect to characteristics other than sex.

Subsequent to these efforts, census technical papers were published analyzing change in the industry and occupation classification systems between 1950 and 1960 (Priebe 1968), between 1960 and 1970 (Priebe et al. 1972), and between 1970 and 1980 (Vines and Priebe 1988). In each case, a subsample of census returns was "double coded." That is, they were coded with the industry and occupation classification for the subsequent census in addition to the one used initially. This provided two kinds of information: the distribution of the labor force for two successive census years coded into the same classification, and a map relating the categories in one classification to the categories in the other. Like the earlier exercises, however, these were very limited. The estimates of industrial and occupational change were at the national level, broken down only by sex. And the mapping between classifications was at an aggregate level, showing for one year the distribution of the categories for the other year that mapped into it. These distributions can tell us nothing about the fate of any particular job. What the mapping does tell us, for example, is that of those classified as working

in "health services, n.e.c." in 1970, 11 percent would have been classified as working in "offices of health practitioners, n.e.c.," 61 percent as working in "health services, n.e.c.," 5 percent as working in "job training and vocational rehabilitation services," and 23 percent as working in "administration of human resources programs," according to the 1980 detailed classification of industries. But the map does not tell us how to assign each 1970 worker in "health services, n.e.c." to a particular 1980 category, which would be necessary, for example, to compare Public-Use Microdata Sample (PUMS) data on individuals from the two censuses.

The inherent difficulty of comparing industry and occupation distributions based on different classification schemes has not deterred analysts from attempting to do so. In the absence of a principled way of converting data from one classification scheme to another, however, analysts have had no choice but to resort to a variety of *ad hoc* matching schemes (e.g., Treiman and Terrell 1975; Williams 1976; Pampel, Land, and Felson 1977; Synder, Hayward, and Hudis 1978; Blau and Hendricks 1979; Rumberger 1981).

A major difficulty with the *ad hoc* schemes commonly employed, apart from their lack of standardization, is that they treat the recalibration process as error free. Thus, inferences about changes over time in industry or occupation characteristics will generally appear stronger than warranted. Moreover, the amount of error is likely to vary substantially, and in unknowable ways, in different parts of the classification scheme. The result is that the analyst of social change is ordinarily hard pressed to know the extent to which observed differences in industrial and occupational data reflect true changes in social structure and the extent to which they represent classification error.

There are only two ways out of this dilemma. One is to return to the original data and recode them with the new classification scheme. In the case of PUMSs from the U.S. census, this is prohibitively expensive. The other is to develop a statistically principled way of converting data from one classification scheme to another, a method that (a) will be relatively accurate and (b) will permit an assessment of the degree of error entailed in the conversion process. In this paper, we evaluate such an effort.

The research reported here is part of an ongoing project evaluating the efficacy of procedures for imputing 1980 codes for detailed industry and occupation into data from the 1970 U.S. census.

The project, which is a cooperative effort involving statisticians, sociologists, and Census Bureau personnel,[2] is the outgrowth of recommendations by the joint Census Bureau/Social Science Research Council Subcommittee on Comparability of Occupation Measurement (1983). The present report is a first evaluation of the accuracy with which industry codes assigned to 1970 data can be recalibrated from the 1970 classification to the 1980 classification. Subsequent reports will present a parallel evaluation of the accuracy with which occupation codes can be converted and will present a number of "worked examples" of procedures for using the recalibrated, or "imputed," data.

The basic approach derives from work on the theory of *multiple imputation*, proposed by Rubin (1978) to handle problems of missing data in surveys and developed in a number of subsequent publications (see, in particular, Rubin 1987, and Rubin and Schenker 1986). The general strategy is to predict, or "impute," the missing values from the relationships existing among variables in those cases without missing data. However, instead of obtaining a single imputation, or point estimate, of the missing values, we repeat the imputation a number of times and create a range of estimates corresponding to the distribution of responses in the complete data. This distribution of estimates—that is, these *multiple imputations*—can then be combined to (a) produce an overall best estimate and (b) compute standard error statistics that reflect both the usual variability inherent in samples and the additional variability due to the imputation process.

In the present case, the complete data are from the 127,125 case Double-Coded Sample (DCS) of 1970 respondents used by the Census Bureau to create the 1980 detailed industry and occupation classifica-

[2] The project is funded by parallel grants from the National Science Foundation to Donald J. Treiman, Department of Sociology, UCLA (SES 83-11483), and Donald B. Rubin, Department of Statistics, Harvard University (SES 83-11428). Clifford Clogg, Departments of Sociology and Statistics, Pennsylvania State University, and William Bielby, Department of Sociology, University of California at Santa Barbara, are consultants to the project. Census Bureau personnel involved in the project include Tom Scopp, Chief, Labor Force Branch, Population Division; John Priebe, Labor Force Branch; Lynn Weidman, Principal Researcher, Statistical Research Division; and Nathaniel Schenker, Undercount Research Staff, Statistical Research Division.

tions. This is a probability sample of labor force cases drawn from responses to the long-form questionnaire administered in 1970 (both the 5 percent and 15 percent versions). The narrative responses to the questions on industry and occupation were assigned codes in the 1980 classification in a special coding operation. Thus, this data set constitutes a sample of the 1970 labor force and includes all the information available in the 1970 PUMS data that was common to the two versions of the long-form questionnaire plus industry and occupation codes in the 1980 classification assigned by trained coders (hereafter referred to as coder-assigned data). (See Priebe [1985] for additional details on the construction of the data set.)

This data set was used to impute, or estimate, five sets of 1980 codes for industry from the respondents' 1970 industry codes and other personal characteristics.[3] (At the time of writing, a similar imputation of 1980 *occupation* codes was not yet completed.) Each 1970 industry category was treated separately. For each 1970 industry, i, a polytomous logistic model was used to represent the probability, p_{ij}, that a person classified in the industry in the 1970 classification scheme would be classified in industry j in the 1980 classification scheme. For estimation purposes we factored this model into a series of conditional dichotomous logistic models that are linear functions of demographic and other personal characteristics of respondents. The characteristics used as predictors were age, race, sex, class of worker, residence in a Standard Metropolitan Statistical Area, education, hours worked per week, weeks worked per year, and region of residence. For each model, the maximum likelihood estimates of the parameters and their covariance matrix were calculated using the appropriate records from the DCS and a specified prior. The posterior distribution of these estimated parameters is multivariate normal. For each imputation, a set of parameters was drawn from these posterior distributions of the conditional models, the probabilities of the 1980 codes were calculated for each record, and an imputed 1980 code was drawn at random according to these probabilities. From a Bayesian perspective, the imputations

[3] The choice of five as the number of imputations is, of course, arbitrary. But it can be shown that the improvement in the accuracy of the imputed data is large for each additional imputation up to about five but then decreases rapidly as the number of imputations increases above five.

are selected from the posterior distribution of the p_{ij}'s. Thus, the set of imputed industry codes for each respondent (in this case, five) corresponds to the distribution of predicted 1980 industry codes for respondents with a given 1970 industry code and a specific set of additional characteristics, i.e., sex, race, age, etc.

Consider (for the sake of simplicity) a hypothetical example. Suppose a 1970 industry, "rice or wheat producers," were subdivided into two categories in 1980: "rice producers" and "wheat producers." Suppose further that rice is grown mainly in Louisiana and Texas and that wheat is grown mainly in the plains states but that there is also substantial wheat production in Texas. Suppose, finally, that in Texas, "rice producers" tend to be black and "wheat producers" tend to be white or Hispanic. Then, a logistic regression equation predicting the odds of being coded in the 1980 classification as a "rice producer" rather than as a "wheat producer" given that one was coded as a "rice or wheat producer" in the 1970 classification would do a good job of distinguishing the two groups, with large parameters for the race and region variables.[4] A 1980 code would then be imputed to each individual by drawing values for the coefficients of each variable from the multivariate normal distributions of the parameters implied by the equation, using these to estimate the conditional odds that the individual would fall into each of the candidate 1980 categories, converting the odds to probabilities, and using the probabilities to assign a 1980 code at random but with probability proportionate to the conditional probabilities. This process would be repeated five times to get five 1980 imputed codes for each individual.

This sort of procedure was applied to each code in the 1970 detailed classification of industries that mapped into more than one

[4] To estimate the actual imputation equations, we dichotomized the race variable ("black" vs. "other") and trichotomized the region variable ("North," consisting of census regions "New England," "Middle Atlantic," and "East North Central"; "South," consisting of census regions "South Atlantic," "East South Central," and "West South Central"; and "West," consisting of census regions "West North Central," "Mountain," and "Pacific"). Residence in the "West" would predict "wheat growing" over "rice growing," and residence in the "South" would predict "rice growing" over "wheat growing." The distinction between "wheat growing" and "rice growing" among Texans (which is in the "South" region) would be represented by the race variable; "black" would predict "rice growing" over "wheat growing."

code in the 1980 classification.[5] The outcome is a set of six 1980 detailed industry codes for each respondent in the DCS: the coder-assigned 1980 code and five imputed codes.[6] These six codes constitute the data for the evaluation exercise in the remainder of this paper.

Two major questions are addressed. First, how well do the imputed codes approximate the coder-assigned ("actual") codes? For these purposes, we treat the coder-assigned codes as "true scores," since they represent the data that would be available if the narrative

[5] Although only 52 of the 213 detailed industry codes for 1970 (excluding "allocation" codes and code 999, "industry not reported") have valid matches with more than one 1980 code (Vines and Priebe 1988, Table 3), we estimated imputation equations for 180 categories because we based our analysis on *uncorrected* 1970 data. The 1970 industry codes we used were those assigned during standard processing of the 1970 census. The 1980 codes were assigned by specially trained coders in the Labor Force Branch of the Census Bureau. In the course of assigning 1980 codes, these coders corrected errors in the 1970 codes whenever they encountered them. The *corrected* 1970 codes were used by Vines and Priebe to create a valid mapping of 1970 codes into 1980 codes. However, since the PUMSs to which we plan to apply the imputations contain *uncorrected* 1970 codes, it seemed to us appropriate to model the relationship between the uncorrected 1970 codes and the 1980 codes in the DCS. The result is, of course, that in some cases we will impute invalid 1980 codes to the 1970 data. However, the alternative would also have led to invalid imputations, but of a much more insidious kind, since the use of the *corrected* 1970 codes in the modeling process would have created imputation equations that implicitly treat erroneous 1970 codes in the PUMS as if they are correct.

To include a 1980 code in the set of codes to which a 1970 code could be mapped, we required that for a given 1970 code, at least two cases exist in the DCS with a common 1980 code. The 1970–1980 matches represented by a single case were not modeled because including them would have introduced excessive noise into the imputed data when imputations were made to the large PUMSs.

In five cases, individuals were assigned 1980 code 990 ("industry not reported") because in these cases, 1970 codes had been erroneously assigned when there was insufficient information to assign a code. Because two of these cases had the same uncorrected 1970 code (609, "department and mail order establishments —retail trade"), 1980 code 990 was regarded as a legitimate code for imputation for this category. As it happens, in nine cases (out of 2,494), at least one of the five imputations assigned code 990. (For imputations 1–3, one case was assigned imputed code 990; for imputation 4, two cases were so assigned; and for imputation 5, three cases were so assigned.) For this reason, the frequencies shown in Table 3 for each of the imputations and the coder-assigned data differ slightly among one another.

[6] For 1970 industries mapping into a single 1980 industry, no imputation equation was estimated. In these cases, each of the "imputed" codes is just the 1980 code into which the 1970 code maps.

descriptions from one or more 1970 PUMSs were directly coded with the 1980 classification. Second, is it worthwhile to use the logistic regression imputation equations to impute 1980 industry codes to two of the 1 percent PUMSs for 1970 (as is currently planned), or would we be better off ignoring the imputations altogether and simply using the DCS with the coder-assigned 1980 scores as the 1970 data set for purposes of comparisons between 1970 and 1980? That is, does the loss in precision from imperfect prediction of 1980 codes more than offset the reduction in sampling variability in the large PUMSs? To anticipate our results, (a) the imputations closely approximate the coder-assigned codes, and (b) it clearly is preferable to impute codes to PUMSs rather than rely on the smaller DCS.

Before beginning our evaluation exercise, we summarize those aspects of the theory of multiple imputation that pertain to our analysis.

2. MULTIPLE-IMPUTATION THEORY

The basic idea underlying the multiple-imputation approach is to partition the uncertainty in statistics based on imputed data—measured by the standard error of these statistics—into two components: a portion associated with variance within each imputation, which is equivalent to the uncertainty associated with sampling variability in the absence of imputation, and a portion associated with the added uncertainty due to imputation. These components are estimated by treating each of the imputations as a variable in a *completed* data set (together with the remaining variables in the data set) and estimating the appropriate complete-data statistics as many times as there are imputations. For example, if we had five imputations (as we do here), we would estimate the statistics of interest five times, using standard complete-data procedures, as if we had five separate data sets. We would then combine these statistics as indicated below, following Rubin (1987, pp. 75–77) and Rubin and Schenker (1986, pp. 366–67), to estimate the components and evaluate them to test hypotheses and establish confidence intervals.

Suppose Q is a statistic to be estimated for which, for complete data, the sampling distribution is approximately normal with mean 0 and variance U. That is, suppose $Q - \hat{Q} \sim N(0, U)$, where $\hat{Q}$ and $\hat{U}$ are estimates of Q and the variance of $Q - \hat{Q}$, respectively. Now, suppose

that $\hat{Q}$ is computed from imputed data and there are m imputations. Each of the data sets containing imputed data is known as a *completed* data set. Let $\hat{Q}_{*i}$ and $\hat{U}_{*i}$ $(i = 1 \ldots m)$ be the estimates of Q and $\hat{U}$ derived from each of the completed data sets. Then it can be shown that the appropriate estimate of Q is

$$\hat{Q}_{*.} = \sum_{i=1}^{m} \hat{Q}_{*i}/m \tag{1}$$

and that the appropriate estimate of the variance of $Q - \hat{Q}_{*.}$ is T_*, given by

$$T_* = \hat{W} + ((m+1)/m)\hat{B}, \tag{2}$$

where

$$\hat{W} = \sum_{i=1}^{m} \hat{U}_{*i}/m \tag{3}$$

is the average within-imputation variance of $(Q - \hat{Q}_{*.})$ and

$$\hat{B} = \sum_{i=1}^{m} \left(\hat{Q}_{*i} - \hat{Q}_{*.}\right)^2/(m-1) \tag{4}$$

is the between-imputation variance of $(Q - \hat{Q}_{*.})$. Inferences about Q can be made by considering the ratio $\hat{Q}_{*.}/T^{1/2}$, which is distributed as t with ν degrees of freedom, where ν is given by

$$\nu = \left[1 + \left(\frac{m}{m+1}\right)\frac{\hat{W}}{\hat{B}}\right]^2 (m-1). \tag{5}$$

These results can be used both to assess how much extra uncertainty is introduced into an analysis by using imputed data rather than complete data and to arrive at correct estimates of uncertainty for an analysis based on imputed data. Specifically, the component $((m+1)/m)\hat{B}$ in equation (2) is a measure of the extra uncertainty introduced by the use of imputed data. Ideally, this component will be small relative to $\hat{W}$, which measures the amount of uncertainty due to sampling variability. Certainly, the larger the added uncertainty due to imputation, the less desirable it is to rely upon imputed data and the greater the need to consider additional data collection or, in the case of the census data, a large-scale recoding effort. The relative size of the two components of equation (2) is, however, an empirical question, which must be considered separately for each specific data set. In the

present case, our interest is in whether recalibration of the 1970 census detailed classification of industries via a multiple-imputation procedure provides sufficiently accurate data to warrant the use of the imputed data for 1970–1980 comparisons. We now turn to an assessment of this question.

3. EVALUATION

Our basic strategy is to compute various statistics of the sort researchers interested in industrial differences in labor force characteristics might compute. We compute parallel statistics using the imputed data and the coder-assigned data and compare them. We report these results as answers to a set of questions a researcher might put to these data. We also consider their implications for the properties of imputed data in a 2 percent PUMS.

How similar are the imputed distributions of industry codes to each other and to the actual 1980 codes? To answer this, we compute an index of dissimilarity, Δ, between each pair of imputations of detailed (three-digit) industry codes and between each set of imputed codes and the actual codes. The index of dissimilarity is defined as

$$\Delta = \frac{\Sigma |p_i - q_i|}{2}, \tag{6}$$

where p_i is the percentage of cases in category i of one distribution, q_i is the percentage of cases in category i of the other distribution, and the summation is over all categories. We interpret Δ in the conventional way, as the percentage of cases in one distribution that would have to be shifted to another category to make the two distributions identical.

Table 1 shows the Δ's for pairs of imputations and for each imputation and the coder-assigned (actual) 1980 industry codes. Thus, in this table, Δ indicates the percentage of cases that would have to be shifted to different categories to make the distributions based on two imputations (or on an imputation and the coder-assigned codes) identical. All the Δ's are very small. The net error in classification resulting from each of the imputations is less than one percent; the average Δ is 0.86. The error in a distribution formed by pooling the five imputations (by averaging over the five percentage distributions) would, of course, be even smaller. (We refer to *net* error of classification, since Δ

TABLE 1

Indexes of Dissimilarity (Δ) Between Five Imputations of 213 1980 Detailed (Three-Digit) Industry Codes and Actual (Coder-Assigned) Detailed Industry Codes, for the Double-Coded Sample

	Imputation				
	2	3	4	5	Actual Codes
Imputation 1	1.04	0.88	0.99	1.01	0.88
Imputation 2		1.03	0.99	0.94	0.87
Imputation 3			1.07	1.10	0.86
Imputation 4				0.92	0.80
Imputation 5					0.89
Mean of Δ's (actual vs. imputed)					0.86

measures the proportion of cases that would have to be shifted among categories to make two distributions identical; Δ does not measure offsetting classification errors, in which person A, who is in category i in the coder-assigned classification, is imputed to category j and person B, who is in category j in the coder-assigned classification, is imputed to category i.)

How accurate are estimates of the dissimilarity in the distribution of population groups across industries based on the imputed data? Table 2 shows Δ's measuring the dissimilarity in the distribution of men and women across detailed industry groups and the dissimilarity in the distribution of blacks and nonblacks, computed separately from each set of imputed codes and also from the coder-assigned codes. Again, the estimates based on the imputed data are very close to those derived from the coder-assigned data. In each case, the Δ's based on the imputed codes are within about one-half percentage point of the Δ computed from the coder-assigned codes.

Note that the Δ's computed from the imputed data tend to be fractionally smaller than the Δ's computed from the coder-assigned data. There is no obvious reason for this, since both sex and race were used as imputation variables. Because the magnitude of the difference is so small, it is of little practical import.

These statistics are quite gross. They tell us that imputation error is not likely to be important when making comparisons across

TABLE 2
Indexes of Dissimilarity (Δ) Between Population Subgroups Across 213 1980 Detailed Industry Codes

Statistics based on	Men vs. Women	Blacks vs. Nonblacks
Imputation 1	44.92	24.74
Imputation 2	45.06	24.69
Imputation 3	45.18	24.57
Imputation 4	44.88	24.20
Imputation 5	45.09	24.38
Mean of Δ's	45.03	24.52
Actual codes	45.48	24.81

industry categories for the entire labor force. This is hardly surprising, since in overall comparisons, between-group differences are likely to be large relative to within-group differences. *Often, however, researchers wish to analyze smaller and more narrowly defined subsamples. What is the magnitude of error in such cases*? To answer this, we assess the error in various statistics computed for each of thirteen major industry groups and for each of eight selected detailed industry categories. The statistics we compute are the mean number of years of school completed by workers in each industry category, the slope coefficient (b) from the regression of annual income on years of school completed, the percentage female, and the percentage black. We also present results for a simple log-linear model relating sex, employment status, and income for two of the thirteen major industry groups.

Since the size of each of the categories is itself determined by the imputation procedure, it is of some interest to know the extent to which the category sizes vary across imputations. Table 3 shows the frequency of each major industry group, based on the coder-assigned codes and each of the imputed codes, and Table 4 shows the corresponding frequencies for the eight detailed industry categories utilized in the subsequent analysis. As expected, the counts vary to some degree, but the variability is very small relative to the size of the categories. Among the thirteen major groups, no imputation yields a count that deviates from that based on the coder-assigned codes by more than 3 percent; and among the eight detailed groups, the largest deviation is

TABLE 3
Frequency Distribution of 1980 Major Industry Groups Based on Coder-Assigned (Actual) Codes and Five Imputations

Major Industry Group	Imputation 1	2	3	4	5	Actual Codes
Agriculture, etc.	4,346	4,334	4,344	4,342	4,330	4,337
Mining	1,006	994	1,014	1,015	1,005	990
Construction	7,846	7,833	7,804	7,877	7,840	7,825
Manufacturing	33,427	33,468	33,479	33,360	33,423	33,492
Transport, etc.	9,584	9,612	9,621	9,590	9,577	9,593
Wholesale trade	5,346	5,348	5,389	5,316	5,389	5,357
Retail trade	19,397	19,337	19,323	19,400	19,347	19,327
Finance, etc.	6,051	6,022	6,039	6,039	6,034	6,007
Business service, etc.	3,955	3,941	3,979	4,029	3,928	3,949
Personal service	5,358	5,402	5,378	5,364	5,390	5,398
Entertainment, etc.	1,055	1,053	1,030	1,038	1,036	1,047
Professional service	20,774	20,765	20,799	20,707	20,800	20,777
Public administration	5,453	5,489	5,399	5,520	5,496	5,495
Total[a]	123,598	123,598	123,598	123,597	123,595	123,594

Note: There are 127,125 cases in the full DCS. But 599 cases assigned "allocation" codes in 1970 were excluded, as were 2,987 cases missing data on one or more of the predictor variables included in the imputation equations. These exclusions yield 123,599 cases for which 1980 industry codes were imputed.

[a] The N's differ from 123,599 because in a few cases that were erroneously coded in the 1970 classification, coders assigned code 990 ("industry not reported") in the 1980 classification. Hence, in a few instances, code 990 was imputed. See note 5 for additional details.

17 percent and the second largest is 11 percent. Both of these deviations are for the smallest industry analyzed here: industry 150 ("miscellaneous textile products"), for which the count based on coder-assigned codes is 101. Thus, it appears that the imputation process produces more or less consistent counts of industries and that the only non-negligible variability in the counts occurs among very small industries, as we expected.

How valid are estimates of the mean years of school completed by workers in each major industry group based on the imputed data? To assess this, consider Tables 5 and 6. Table 5 shows the mean years of school completed by workers in each major industry group estimated from each of the five imputations, the mean of these five estimates, the mean

TABLE 4
Frequency Distribution of Selected 1980 Detailed Industry Groups Based on Coder-Assigned (Actual) Codes and Five Imputations

Major Industry Group	Imputation					Actual Codes
	1	2	3	4	5	
030 Forestry	109	109	116	112	111	107
150 Miscellaneous textile products	103	84	99	90	94	101
331 Machinery, except electrical	1,545	1,570	1,515	1,540	1,566	1,529
470 Water and irrigation	237	234	241	234	233	232
601 Grocery stores	2,734	2,705	2,671	2,691	2,692	2,685
712 Real estate	1,213	1,199	1,198	1,229	1,214	1,189
741 Detective services	140	135	143	138	142	136
860 Education services	194	171	167	199	193	183

years of school completed estimated from the coder-assigned data, and the difference between the estimates based on the coder-assigned data and the imputed data. When we compare the mean of the estimates based on the imputed data and the estimate based on the coder-assigned data, we see that they are very close: The largest discrepancy is four points in the second decimal place. If these estimates were based on independent samples, we would conclude that the samples were drawn from the same population, because when the standard t test of the significance of the difference between two means is applied, none of the t ratios exceeds unity. Moreover, the estimates based on the coder-assigned data are neither systematically larger nor smaller than the mean of the estimates based on the imputed data: Six are larger, five are smaller, and two are identical to two decimal places. Once again, substantive conclusions would not be affected by use of the mean of the imputed estimates rather than estimates based on the coder-assigned data.

What is the effect of the imputation procedure on the reliability of the estimates of mean years of school completed by workers in each industry? That is, what is the magnitude of the standard errors of means computed from the imputed data relative to the standard errors of means computed from the coder-assigned data? This information is presented in Table

TABLE 5

Mean Years of School Completed by Workers in Thirteen Major Industry Groups, Estimated from Five Imputed Distributions and Actual Distribution of 1980 Detailed Industry Codes

Major Industry Group	Imputation 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	Mean (6)	Actual Data (7)	Difference: (7)–(6) (8)
Agriculture, etc.	9.76	9.76	9.78	9.77	9.78	9.77	9.74	−0.03
Mining	10.93	10.93	10.91	10.91	10.94	10.92	10.96	0.04
Construction	10.37	10.36	10.37	10.37	10.37	10.37	10.34	−0.03
Manufacturing	10.99	10.99	10.99	10.99	10.99	10.99	10.98	−0.01
Transport, etc.	11.33	11.34	11.34	11.33	11.33	11.33	11.33	0.00
Wholesale trade	11.53	11.57	11.55	11.56	11.57	11.56	11.58	0.02
Retail trade	11.13	11.13	11.14	11.13	11.13	11.13	11.13	0.00
Finance, etc.	12.58	12.58	12.58	12.58	12.59	12.58	12.60	0.02
Business service, etc.	11.72	11.67	11.69	11.72	11.69	11.70	11.67	−0.03
Personal service	10.01	10.03	10.04	10.02	10.03	10.03	9.99	−0.04
Entertainment, etc.	11.74	11.64	11.64	11.64	11.64	11.66	11.67	0.01
Professional service	13.67	13.67	13.66	13.68	13.66	13.67	13.69	0.02
Public administration	12.43	12.47	12.44	12.43	12.44	12.44	12.47	0.03

6. Columns 1 and 2 show the standard errors of the means computed from the imputed data (following equations [2]–[4]) and the means computed from the coder-assigned, or actual, data, and column 10 shows their ratio. When we inspect the ratios, we see that for most major industry groups, the increase in the standard error due to imputation is negligible.

For four major industry groups ("wholesale trade," "business and repair services," "entertainment and recreation services," and "public administration"), however, the increase in the standard errors is large enough to take seriously, ranging from 12 to 16 percent. It is instructive to consider the sources of this increase. As we noted above, the standard error of a statistic based on the imputed data has two sources: within-imputation sampling variance and between-imputation variance (equations [3] and [4]). Column 4 shows the within-imputa-

TABLE 6

Estimates of Standard Error of Mean Years of School Completed for Thirteen Major Industry Groups, Actual and Imputed Data

Major Industry Group	SE of Actual Data (1)	SE of Imputed Data (2)	Variance of Imputed Data (3)	Within-Imputation Sampling Variance (4)	Adjusted Between-Imputation Variance[a] (5)	Expected SE for PUMS[b] (6)	SE for PUMS Without Imputation[b] (7)	Ratio: (1)/(6) (8)	Ratio: (6)/(7) (9)	Ratio: (2)/(1) (10)
Agriculture, etc.	0.04905	0.05009	0.00251	0.00239	0.00012	0.01779	0.01401	2.76	1.27	1.02
Mining	0.10075	0.10129	0.01026	0.01004	0.00022	0.03227	0.02873	3.12	1.12	1.01
Construction	0.03233	0.03260	0.00106	0.00104	0.00002	0.01046	0.00924	3.09	1.13	1.01
Manufacturing	0.01557	0.01563	0.00024	0.00024	0.00000	0.00448	0.00448	3.48	1.00	1.00
Transport, etc.	0.02542	0.02620	0.00069	0.00065	0.00004	0.00946	0.00731	2.69	1.29	1.03
Wholesale trade	0.03662	0.04149	0.00172	0.00139	0.00034	0.02121	0.01067	1.73	1.99	1.13
Retail trade	0.01712	0.01786	0.00032	0.00030	0.00002	0.00695	0.00492	2.46	1.41	1.04
Finance, etc.	0.03122	0.03150	0.00099	0.00097	0.00002	0.01018	0.00892	3.07	1.14	1.01
Business service, etc.	0.04535	0.05085	0.00259	0.00202	0.00056	0.02702	0.01289	1.68	2.10	1.12
Personal service	0.03852	0.04079	0.00166	0.00151	0.00016	0.01673	0.01113	2.30	1.50	1.06
Entertainment, etc.	0.08437	0.09789	0.00958	0.00718	0.00240	0.05468	0.02430	1.54	2.25	1.16
Professional service	0.02234	0.02420	0.00059	0.00050	0.00008	0.01119	0.00642	2.00	1.74	1.08
Public administration	0.03534	0.04002	0.00160	0.00128	0.00032	0.02071	0.01025	1.71	2.02	1.13

[a]Adjusted by ratio $((m + 1)/m)$. See equation (2).

[b]See text for explanation.

tion variance and column 5 shows the adjusted between-imputation variance; column 3 shows their sum (equation [2]). Column 2 is, of course, just the square root of column 3. For all thirteen major industry groups, the within-imputation variance is large relative to the adjusted between-imputation variance,[7] but the smallest ratios are those for the four industries with non-negligible imputation error. Columns 4 and 5 show that this is due to the relatively large size of the between-imputation variance and not to anything systematic about the within-imputation variance. For whatever reasons, the imputation procedure assigned detailed 1980 industry codes in such a way as to produce in these groups a substantial amount of variability from one imputation to another in mean years of schooling. This, of course, is just another way of saying that education is not a very good predictor of whether a particular individual will be included in each of these four 1980 major industry groups, conditional on inclusion in particular 1970 industry categories.

From a practical standpoint, the important question is not the size of the standard errors in the imputed data for the DCS, but rather the size of the standard errors we can expect in the 2 percent PUMS into which 1980 industry codes will be imputed. Put simply, *the issue is whether a researcher would be better off using the coder-assigned data for the DCS or the imputed data in the 2 percent PUMS*. The data necessary to answer this question are shown in columns 6 and 8 of Table 6. Column 6 shows estimates of the standard errors we expect in the PUMS, given that the PUMS is simply a larger sample drawn from the same population as the DCS. (Actually, this is not strictly correct, since the DCS is clustered by enumeration districts [Priebe 1985] and hence is not as efficient as a random sample. For this reason, our estimates of the improvement in the standard errors in the PUMS relative to the DCS are conservative. Unfortunately, estimates of the size of the design effect in the DCS are not available, so we cannot derive more precise estimates of the extent of improvement.)

The estimates in column 6 were derived by assuming that the between-imputation variance in the PUMS will be the same as in the

[7] The between-imputation variance is adjusted by $(m+1)/m = 6/5$ to take account of the fact that the number of imputations is finite, as per equation (2).

DCS but that the within-imputation variance will be reduced by a factor of 12.168, which is the ratio of the size of 2 percent of the labor force in 1970 (U.S. Bureau of the Census 1973, Table 1) to the number of cases for which industry was imputed in the DCS.[8] This is, in fact, a very conservative assumption, since, in general, the between-imputation variance should also decrease as sample size increases.[9]

Column 8 shows that the expected standard errors in the PUMS are always substantially smaller than the standard errors estimated from coder-assigned data in the DCS. Thus, insofar as these results hold for other variables and population subgroups, we can conclude

[8] The size of the labor force in 1970 was 80,071,130 (U.S. Bureau of the Census 1973, Table 1). However, 103,340 persons were unemployed and had not worked since 1959; no information on industry was available for these persons. For the remainder of the labor force, those not providing sufficient information to permit the coding of industry were allocated to industry major groups and were assigned special "allocation" codes in the 1970 industry classification. Although these categories will be imputed in the PUMS, they were not used in the exercise reported here. Hence, the 4,771,640 persons in the 1970 labor force with allocation codes for industry were excluded from the base used to calculate the denominator of the ratio of the size of the card-deck sample to the size of the 2 percent PUMS. The resulting base is 75,196,150 (= 80,071,130 − 103,340 − 4,771,640), of which 2 percent is 1,503,923. Similarly, we excluded from the DCS 539 allocation cases and 2,987 cases missing data on the independent variables used in the logistic regression models, leaving an effective sample of 123,599. (Note that there are no missing data in the PUMSs, since all missing values are imputed as part of the editing process; see U.S. Bureau of the Census [1976, pp. 15–65, 15–66] for a discussion of the procedures utilized.) The ratio 1,503,923/123,599 = 12.168.

[9] Our computations rely on comparisons of actual data from the DCS with imputations to the same data set. Multiple imputation of the PUMS would, of course, apply logistic results from the DCS to new data. However, in both instances, logistic coefficients for each of the five imputations are sampled from a multivariate normal distribution with a covariance matrix estimated from the logistic regression. Thus, sampling variability in the logistic coefficients is represented in the imputation procedure, even when it is applied to the data that generated the coefficients. Consequently, the coefficients do not "overfit" the DCS, and the between-imputation variance should be the same in both data sets except insofar as it is affected by sample size. According to Rubin (pers. comm. 1987), the ratio of between-imputation variance to total (within- plus between-imputation) variance should not be affected by sample size, which suggests that the between-imputation variance as well as the within-imputation variance would be reduced by a factor of 12.168 in the 2 percent PUMS. We have, however, chosen to make an extremely conservative assumption regarding the between-imputation variance to avoid any chance of overstating the advantage accruing from use of the imputed data.

that there is a real gain in precision when we use imputed data in the PUMS rather than coder-assigned data in the smaller DCS.

It is also important to assess the magnitude of the increase in error resulting from the use of imputed data in the PUMS relative to the error we would expect in PUMS data from sampling variability alone. Column 7 shows the hypothetical standard errors that we would expect if we had double-coded the PUMS. They are, in fact, the same as the estimates in column 6 except that the between-imputation variance is assumed to be zero. Column 9 shows the ratio of the two PUMS estimates and thus gives us a hypothetical estimate of the cost of imputation in samples as large as the 2 percent PUMS. When we inspect the coefficients in column 9, we see that for one industry, "manufacturing," there is no error associated with imputation (to the level of precision associated with two decimal places) but that for many industries, the increase in the standard errors is substantial. Moreover, as before, the effect of imputation on the standard errors varies substantially across industries, increasing the standard error by a factor of 2.25 for the most strongly affected industry, "entertainment and recreation services." From these results, we conclude that a researcher could err seriously by using the imputed data as if they were double-coded data and neglecting to compute the correct standard errors. Not only would this lead to the conclusion that the data are more precise than they actually are, but the relative precision of different statistics would change in unknown ways.

So far, we have focused on comparisons of major industry groups and thus have dealt with samples of about 1,000 cases or larger. We now study the same statistic, mean years of school completed, for eight selected detailed industry categories. These subsamples are, of course, much smaller; there are only 101 cases in the smallest, "miscellaneous textile products." Tables 7 and 8 show data corresponding to the data in Tables 5 and 6. Generally, the conclusions we reach are similar to those we discussed above. The means based on the imputed data are similar to those based on the coder-assigned data, and the largest discrepancy is 0.06 (again, no t ratio for the significance of the difference between two means exceeds unity); there are both positive and negative deviations of the imputed means from the means based on the coder-assigned data; the expected reductions in the standard errors in the PUMS relative to the standard errors for the coder-assigned data in the DCS are on the whole large; and the increases in the

TABLE 7

Mean Years of School Completed by Workers in Eight Detailed (Three-Digit) Industry Categories, Estimated from Five Imputed Distributions and Actual Distribution of 1980 Detailed Industry Codes

Detailed Industry Category	Imputation 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	Mean (6)	Actual Data (7)	Difference: (7)–(6) (8)
030 Forestry	12.33	12.39	12.40	12.38	12.50	12.40	12.46	0.06
150 Miscellaneous textile products	10.11	9.94	10.21	10.44	10.48	10.24	10.19	−0.05
331 Machinery, except electrical	11.16	11.21	11.21	11.21	11.20	11.20	11.19	−0.01
470 Water and irrigation	10.76	10.85	10.78	10.76	10.84	10.80	10.78	−0.02
601 Grocery stores	10.94	10.93	10.94	10.94	10.92	10.93	10.93	0.00
712 Real estate	11.68	11.73	11.73	11.75	11.77	11.73	11.75	0.02
741 Detective services	10.86	10.93	10.91	10.90	10.93	10.91	10.96	0.05
860 Educational services	13.68	13.67	13.40	13.68	13.51	13.59	13.63	0.04

TABLE 8

Estimates of Standard Error of Mean Years of School Completed for Eight Detailed (Three-Digit) Industry Categories, Actual and Imputed Data

Detailed Industry Category	SE of Actual Data (1)	SE of Imputed Data (2)	Variance of Imputed Data (3)	Within-Imputation Sampling Variance (4)	Adjusted Between-Imputation Variance (5)	Expected SE for PUMS (6)	SE for PUMS Without Imputation (7)	Ratio: (1)/(6) (8)	Ratio: (6)/(7) (9)	Ratio: (2)/(1) (10)
030 Forestry	0.32869	0.31708	0.10054	0.09592	0.00462	0.11182	0.08879	2.94	1.26	0.96
150 Miscellaneous textile products	0.34826	0.45226	0.20454	0.14295	0.06160	0.27082	0.10839	1.29	2.50	1.30
331 Machinery, except electrical	0.06624	0.07110	0.00506	0.00449	0.00056	0.03055	0.01921	2.17	1.59	1.07
470 Water and irrigation	0.17923	0.18946	0.03590	0.03359	0.00230	0.07117	0.05254	2.52	1.35	1.06
601 Grocery stores	0.04188	0.04328	0.00187	0.00178	0.00010	0.01556	0.01209	2.69	1.29	1.03
712 Real estate	0.09193	0.09661	0.00933	0.00807	0.00126	0.04386	0.02576	2.10	1.70	1.05
741 Detective services	0.21952	0.22676	0.05142	0.05042	0.00100	0.07169	0.06437	3.06	1.11	1.03
860 Educational services	0.17815	0.24921	0.06210	0.04258	0.01952	0.15173	0.05916	1.17	2.57	1.40

standard errors in the PUMS relative to the size of the standard errors we would expect if the 1980 industry codes in the PUMS were based on coder-assigned rather than imputed data are also non-negligible.

One figure in Table 8 requires special comment. For workers in "forestry," the standard error based on the imputed data is *smaller* than the standard error based on the coder-assigned data. This can arise when the within-imputation sampling variance is smaller than the sampling variance of the mean computed from the coder-assigned data, as it is in the current case. In turn, this implies that the average standard deviation of education is smaller in the imputed data than in the coder-assigned data, since the frequencies in the imputed and coder-assigned data are about the same size (3.27 for the imputed data and 3.40 for the coder-assigned data). Were this to arise frequently or systematically, it would imply that our imputation procedure was overdetermining the relationship between respondents' 1970 characteristics and their 1980 industry codes. However, of the 45 comparisons in Tables 6, 8, 10, 12, and 14, only four show smaller standard errors for statistics computed from the imputed data than for statistics computed from the coder-assigned data, and for the most part they are only trivially smaller.

Tables 9 through 14 replicate the analysis presented above for three additional variables: the slope coefficient from the regression of annual income on years of school completed, the percentage female, and the percentage black. In each case, comparisons are made for the eight detailed industry categories. Because the results have been so consistent, we felt it unnecessary to report the corresponding data for the thirteen major industry groups, which all involve much larger sample sizes. We have made the computations and inspected them. They contain no surprises and hence we do not include them here.

There is little news in Tables 9 through 14. Again, a detailed analysis of these tables would yield the same conclusions that we have already drawn. But two additional points may be noted. First, the relative size of the statistics based on the actual and imputed data in each of the industries is not consistent across statistics. Thus, the value of one statistic based on the coder-assigned data may be larger than that of the corresponding statistic based on the imputed data, while the relative sizes of a different statistic computed for the coder-assigned and imputed data may be reversed. Similarly, not only does the relative importance of sampling error and imputation error as compo-

TABLE 9

Slope Coefficients (b's) from Regressions of Annual Income on Years of School Completed by Workers in Eight Detailed (Three-Digit) Industry Categories, Estimated from Five Imputed Distributions and Actual Distribution of 1980 Detailed Industry Codes

Detailed Industry Category	Imputation					Mean	Actual Data	Difference: (7)–(6)
	1	2	3	4	5			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
030 Forestry	691	746	679	685	824	725	755	30
150 Miscellaneous textile products	697	515	652	634	738	647	631	−16
331 Machinery, except electrical	634	615	624	644	637	631	527	−104
470 Water and irrigation	500	381	403	393	499	435	348	−87
601 Grocery stores	326	309	321	309	294	312	297	−15
712 Real estate	714	705	700	738	702	712	681	−31
741 Detective services	100	122	136	120	126	121	143	22
860 Educational services	311	202	27	321	263	225	36	−189

TABLE 10

Estimates of Standard Error of b (Slope Coefficient for Regression of Annual Income on Years of School Completed) for Eight Detailed (Three-Digit) Industry Categories, Actual and Imputed Data

Detailed Industry Category	SE of Actual Data (1)	SE of Imputed Data (2)	Variance of Imputed Data (3)	Within-Imputation Sampling Variance (4)	Adjusted Between-Imputation Variance (5)	Expected SE for PUMS (6)	SE for PUMS Without Imputation (7)	Ratio: (1)/(6) (8)	Ratio: (6)/(7) (9)	Ratio: (2)/(1) (10)
030 Forestry	136	154	23,583	19,049	4,534	78	40	1.74	1.97	1.13
150 Miscellaneous textile products	132	167	27,949	19,430	8,520	101	40	1.31	2.52	1.27
331 Machinery, except electrical	57	60	3,616	3,460	156	21	17	2.72	1.24	1.05
470 Water and irrigation	91	120	14,369	10,162	4,207	71	29	1.28	2.46	1.32
601 Grocery stores	41	43	1,834	1,649	186	18	12	2.29	1.54	1.04
712 Real estate	70	72	5,220	4,928	292	26	20	2.65	1.31	1.03
741 Detective services	119	117	13,773	13,565	208	36	33	3.27	1.09	0.99
860 Educational services	138	183	33,400	16,064	17,337	136	36	1.01	3.76	1.32

TABLE 11

Percentage Female Among Workers in Eight Detailed (Three-Digit) Industry Categories, Estimated from Five Imputed Distributions and Actual Distribution of 1980 Detailed Industry Codes

Detailed Industry Category	Imputation 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	Mean (6)	Actual Data (7)	Difference: (7)–(6) (8)
030 Forestry	9.2	11.0	10.3	14.3	12.6	11.5	10.3	−1.2
150 Miscellaneous textile products	35.0	35.7	32.3	33.3	33.0	33.9	33.7	−0.2
331 Machinery, except electrical	17.2	16.8	18.5	18.0	18.5	17.8	17.5	−0.3
470 Water and irrigation	14.8	14.5	14.9	14.5	16.3	15.0	15.1	0.1
601 Grocery stores	38.3	38.7	38.5	37.9	37.8	38.2	38.1	−0.1
712 Real estate	38.9	39.8	38.5	38.9	38.7	39.0	38.4	−0.6
741 Detective services	12.1	12.6	13.3	11.6	12.7	12.5	13.2	0.7
860 Educational services	67.0	70.8	72.5	71.9	64.2	69.3	68.9	−0.4

nents of the total standard error vary across industries for each statistic, but the pattern of industry variability varies from statistic to statistic. These results are encouraging because they suggest that the imputation procedure has introduced no systematic distortion pertaining to identifiable subsamples—at least for the limited number of comparisons we have made here. The results also underscore the importance of explicitly computing the impact of imputation unreliability on statistical precision for *all* statistics, since the ratio of sampling to imputation error varies across statistics.

Second, for one industry, "educational services, n.e.c.," the standard error expected from the 2 percent PUMS is *larger* than the standard error computed from the coder-assigned data in the DCS for two of the four statistics studied. This serves to remind us that imputation is not without cost. When the imputation is very uncertain, as it apparently is in the assignment of cases to "educational services, n.e.c.," there is a net loss in precision when the imputed codes in the

TABLE 12

Estimates of Standard Error of Percentage Female Among Workers in Eight Detailed (Three-Digit) Industry Categories, Actual and Imputed Data

Detailed Industry Category	SE of Actual Data (1)	SE of Imputed Data (2)	Variance of Imputed Data (3)	Within-Imputation Sampling Variance (4)	Adjusted Between-Imputation Variance (5)	Expected SE for PUMS (6)	SE for PUMS Without Imputation (7)	Ratio: (1)/(6) (8)	Ratio: (6)/(7) (9)	Ratio: (2)/(1) (10)
030 Forestry	2.94	3.73	13.90	9.09	4.81	2.36	0.86	1.25	2.73	1.27
150 Miscellaneous textile products	4.70	5.14	26.39	23.94	2.45	2.10	1.40	2.24	1.50	1.09
331 Machinery, except electrical	0.97	1.29	1.66	0.95	0.71	0.89	0.28	1.09	3.19	1.33
470 Water and irrigation	2.35	2.47	6.08	5.41	0.67	1.06	0.67	2.22	1.59	1.05
601 Grocery stores	0.94	1.03	1.05	0.88	0.18	0.50	0.27	1.88	1.86	1.09
712 Real estate	1.41	1.50	2.26	1.96	0.30	0.68	0.40	2.08	1.69	1.07
741 Detective services	2.90	2.88	8.31	7.81	0.50	1.07	0.80	2.72	1.33	0.99
860 Educational services	3.42	5.16	26.66	11.50	15.16	4.01	0.97	0.85	4.13	1.51

TABLE 13

Percentage Black Among Workers in Eight Detailed (Three-Digit) Industry Categories, Estimated from Five Imputed Distributions and Actual Distribution of 1980 Detailed Industry Codes

Detailed Industry Category	Imputation 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	Mean (6)	Actual Data (7)	Difference: (7)–(6) (8)
030 Forestry	6.4	4.6	4.3	4.5	4.5	4.9	5.6	0.7
150 Miscellaneous textile products	10.7	15.5	12.1	15.6	16.0	14.0	13.9	−0.1
331 Machinery, except electrical	4.6	3.9	4.6	4.7	4.6	4.5	4.3	−0.2
470 Water and irrigation	8.0	8.1	7.9	9.0	7.7	8.1	7.8	−0.3
601 Grocery stores	5.6	5.4	5.4	5.7	5.8	5.6	5.6	0
712 Real estate	6.9	8.0	7.0	7.1	6.9	7.2	7.3	0.1
741 Detective services	5.0	5.2	4.9	5.1	4.9	5.0	5.9	0.9
860 Educational services	4.1	5.8	8.4	4.5	5.7	5.7	4.4	−1.3

large PUMS are used rather than the coder-assigned codes in the DCS. Fortunately, however, this is likely to occur only infrequently, because in only 2 of the 45 comparisons made here are the expected standard errors for the large PUMS larger than those based on the actual data in the DCS. Hence, we think that in most situations, analysts would be better served by the large PUMS than by the smaller DCS.

How valid are inferences about log-linear models made from imputed data? Thus far we have dealt with scalar variables—means, proportions, and regression coefficients. But many problems require the analysis of categorical data, for example, via log-linear modeling procedures. Hence, it is useful to assess the adequacy of imputed data for these sorts of problems as well. Typically, log-linear analysis involves first finding the best fitting model from among a set of hierarchical models and then estimating and analyzing the effect parameters associated with that model. The effect parameters constitute no special problem. We simply treat them the same way we treat any other scalar statistic, that is, the same way we dealt with the means, proportions, and regression coefficients above. Finding the best fitting model is, however, a bit more cumbersome. To do this, we need to generate for each

TABLE 14

Estimates of Standard Error of Percentage Black Among Workers in Eight Detailed (Three-Digit) Industry Categories, Actual and Imputed Data

Detailed Industry Category	SE of Actual Data (1)	SE of Imputed Data (2)	Variance of Imputed Data (3)	Within-Imputation Sampling Variance (4)	Adjusted Between-Imputation Variance (5)	Expected SE for PUMS (6)	SE for PUMS Without Imputation (7)	Ratio: (1)/(6) (8)	Ratio: (6)/(7) (9)	Ratio: (2)/(1) (10)
030 Forestry	2.22	2.25	5.06	4.16	0.90	1.12	0.58	1.99	1.91	1.01
150 Miscellaneous textile products	3.44	4.46	19.90	12.91	6.99	2.84	1.03	1.21	2.76	1.30
331 Machinery, except electrical	0.52	0.64	0.41	0.28	0.13	0.39	0.15	1.33	2.58	1.23
470 Water and irrigation	1.76	1.86	3.47	3.17	0.30	0.75	0.51	2.34	1.47	1.06
601 Grocery stores	0.44	0.48	0.23	0.20	0.04	0.23	0.13	1.90	1.84	1.09
712 Real estate	0.75	0.90	0.81	0.55	0.26	0.55	0.21	1.36	2.60	1.19
741 Detective services	2.02	1.85	3.44	3.42	0.02	0.55	0.53	3.68	1.04	0.92
860 Educational services	1.52	2.52	6.34	2.95	3.39	1.91	0.49	0.80	3.87	1.66

candidate model an adjusted likelihood ratio χ^2 statistic (hereafter, L^2) that reflects the additional uncertainty due to imputation. As before, we begin with the *complete-data* L^2's estimated for each of the completed (imputed) data sets and combine them to obtain the adjusted L^2. Specifically (following Rubin 1987, pp. 99–102), we compute for a given model

$$\hat{D}_m = \frac{\dfrac{\bar{d}_m}{k} - \dfrac{m-1}{m+1}\hat{r}_m}{1+\hat{r}_m}, \tag{7}$$

where $\bar{d}_m$ is the mean of the L^2 values estimated from each of the five sets of imputed data; $m = 5$, the number of imputations; k is the number of degrees of freedom associated with a given hypothesis, that is, a given L^2; and $\hat{r}_m$ is given by

$$\hat{r}_m = \frac{(1+1/m)s_d^2}{2\bar{d}_m + \left[4\bar{d}_m^2 - 2ks_d^2\right]_+^{1/2}}, \tag{8}$$

where s_d^2, the variance in the L^2's, is

$$s_d^2 = \sum_{i=1}^{m} \left(d_i - \bar{d}_m\right)^2/(m-1), \tag{9}$$

and the "+" associated with the bracketed quantity in equation (8) indicates that the quantity must be non-negative. Therefore, computed negative values (which sometimes occur) are set to zero.

The statistic $\hat{D}_m$ is distributed as F with k and $(1+1/k)\hat{\nu}/2$ degrees of freedom, where $\hat{\nu}$ is estimated by

$$\hat{\nu} = (m-1)(1+1/\hat{r}_m)^2. \tag{10}$$

The statistic $\hat{D}_m$ has a one-to-one correspondence to L^2 with k degrees of freedom; the correspondence is defined by the values of $\hat{D}_m$ and L^2 associated with a given probability level on the F and χ^2 distributions, respectively. Thus, to find the L^2 value corresponding to a $\hat{D}_m$ value, we determine the p value associated with the value of $\hat{D}_m$, from an F table, and then, in a χ^2 table, find the L^2 value corresponding to that p value.[10] We convert $\hat{D}_m$ to L^2 rather than make inferences directly

[10] In practice, this is best done by using an algorithm for converting F values to χ^2 values (an inverse χ^2 function), since tables of F values give too crude an approximation. We are indebted to T. E. Raghunathan, Department of Statistics, Harvard University, and to Nathaniel Schenker, U.S. Bureau of the Census, for their assistance in effecting the conversions reported here.

TABLE 15

Likelihood Ratio χ^2 (L^2) Statistics for Log-Linear Models of the Relationship between Sex (S), Employment Status (E), and Earnings (I) for Two Major Industry Groups ("Construction" and "Entertainment and Recreation Services"), Estimated from Actual and Imputed Data

Model	*df*	Imputation 1	2	3	4	5	Based on Imputed Data	Based on Actual Data	Probability (p) Imputed Data	Probability (p) Actual Data
Entertainment										
(1) $[S][E][I]$	12	122.21	129.73	127.37	131.32	129.74	54.50	123.96	$p < 0.0000005$	[b]
(2) $[SE][I]$	8	109.06	113.82	106.50	117.43	115.22	[a]	109.24	[b]	[b]
(3) $[SI][E]$	10	22.32	23.89	29.12	23.09	27.08	21.64	23.73	$0.02 > p > 0.01$	$0.01 > p > 0.005$
(4) $[EI][S]$	10	116.51	120.75	124.13	122.73	119.16	49.86	117.76	$p < 0.0000005$	[b]
(5) $[SE][SI]$	6	9.16	7.98	8.25	9.20	12.56	8.03	9.01	$0.25 > p > 0.10$	$0.25 > p > 0.10$
(6) $[SE][EI]$	6	103.36	104.84	103.26	108.85	104.65	[a]	103.04	[b]	[b]
(7) $[SI][EI]$	8	16.62	14.90	25.88	14.50	16.51	10.30	17.53	$0.25 > p > 0.10$	$0.05 > p > 0.02$
(8) $[SE][SI][EI]$	4	3.16	0.58	5.85	1.81	2.05	0.17	3.67	$p > 0.995$	$0.5 > p > 0.25$
(5)–(8)	2	6.00	7.40	2.13	7.39	10.51	4.17	5.43	$0.25 > p > 0.10$	$0.1 > p > 0.05$
(7)–(8)	4	13.46	14.32	20.03	12.69	14.46	11.88	15.41	$0.02 > p > 0.01$	$0.005 > p > 0.001$
N		793	784	775	783	780	—	781	—	—

Construction										
(1) $[S][E][I]$	12	419.51	426.93	410.58	412.02	394.40	[a]	380.23	[b]	[b]
(2) $[SE][I]$	8	290.74	284.82	283.16	281.33	274.40	[a]	259.71	[b]	[b]
(3) $[SI][E]$	10	146.56	159.33	146.74	147.28	137.87	[a]	137.23	[b]	[b]
(4) $[EI][S]$	10	400.44	406.01	388.87	391.69	374.43	[a]	362.07	[b]	[b]
(5) $[SE][SI]$	6	17.79	17.22	19.32	16.59	17.88	17.36	16.71	$0.01 > p > 0.005$	$0.02 > p > 0.01$
(6) $[SE][EI]$	6	271.67	263.91	261.44	260.99	254.43	[a]	241.55	[b]	[b]
(7) $[SI][EI]$	8	127.49	138.41	125.03	126.95	117.91	[a]	119.07	[b]	[b]
(8) $[SE][SI][EI]$	4	1.78	1.42	1.53	0.72	1.65	1.28	1.87	$0.9 > p > 0.75$	$0.9 > p > 0.75$
(5)–(8)	2	16.01	15.80	17.79	15.87	16.23	16.67	14.84	$p < 0.0005$	$p < 0.001$
(7)–(8)	4	125.71	136.99	123.50	126.23	116.26	[a]	117.20	[b]	[b]
N		7,413	7,384	7,362	7,436	7,393	—	7,387	—	—

[a] χ^2 is very large. χ^2 cannot be computed by the available inverse χ^2 function.
[b] $p < 0.00000005$.

from the values of $\hat{D}_m$ because F statistics are not additive when comparing nested models; χ^2 statistics are.[11]

To assess the adequacy of imputed data for fitting log-linear models, we estimate for two major industry groups ("entertainment and recreation services" and "construction") a simple log-linear model relating income (three categories), class of worker (self-employed, salaried, and government worker), and sex. Table 15 shows the results. Several points may be noted. First, and most important, on the whole, the L^2 values computed from the imputed data lead to the same inferences as the L^2 values computed from the coder-assigned (actual) data. For example, using the conventional 0.05 significance level, we find only one instance in the table in which the actual and imputed data would lead to a different inference: the contrast of models 7 and 8 for "entertainment and recreation services." Using a 0.01 level of significance would lead to a different inference in three cases, but not all in the same direction: The imputed data (but not the actual data) would lead us to conclude that model 5 does not fit the data for "construction"; and the actual data (but not the imputed data) would lead us to conclude that model 5 does not fit the data for "entertainment and recreation services" and that the fit of model 8 is significantly better than the fit of model 7 (also for "entertainment and recreation services"). Second, the L^2's computed from the imputed data are neither uniformly larger nor uniformly smaller than the L^2's computed from the actual data. For "entertainment and recreation services," the L^2's estimated from the combined imputed data (using the inverse χ^2 function) are all smaller than those estimated from the actual data, but this is not true for "construction." Nor is there a consistent difference between the actual and imputed L^2's for contrasts between models. In sum, these results reinforce our earlier conclusions based on scalar data: Imputed data yield results generally similar to those obtained from coder-assigned data, but it is necessary in each instance to carry out the computations required to get correct adjusted L^2 values from the imputed data.

Inferences about differences between two statistics. Using imputed data to make inferences about differences between two statistics, such as means, proportions, or effect parameters for log-linear models, is a

[11] Rubin (1987) also presents other methods for obtaining p values that are more accurate but require more effort.

straightforward application of the procedures discussed above. The statistic in question—e.g., the difference between two means—is computed for each of the completed data sets, and the five estimates are averaged, as per equation (1). The standard error of the statistic is also computed for each of the completed data sets and the five estimates are combined as per equations (2)–(4). The ratio of the statistic estimated by equation (1) to its standard error (the square root of the quantity estimated by equation [2]) is then computed and evaluated with reference to a t distribution, for degrees of freedom estimated as per equation (5).

CONCLUSION

Although the theory of multiple imputation is well developed, there have been few applications of multiple-imputation procedures to large-scale practical problems. Indeed, we believe that the Census project is the first such application. Not only is the problem of recalibrating the industry (and occupation) codes assigned to 1970 data to the 1980 classification of great practical importance to students of social change, but it also constitutes an important test of the efficacy of multiple imputation as a practical solution to recalibration problems of this sort. The complexity of the application, involving hundreds of logistic regression equations and the imputation of a large number of distinct industry codes, meant that the adequacy of the imputations could not be assessed analytically but had to be evaluated by an empirical exercise—specifically, by comparing the results obtained from imputed data with those obtained from the coder-assigned data that the imputations were intended to approximate.

The conclusion of this first exercise, involving only the industry imputations, is that multiple imputation is a viable approach to the recalibration of detailed census classifications. It yields results closely approximating those obtainable from data that have been directly recoded from the narrative reports of respondents, and it provides a way of recalibrating full PUMSs at a small fraction of the cost required for direct recoding.

However, these results also suggest that the appropriate use of imputed data for analysis will require somewhat more complex and extensive statistical manipulations than analysts are used to. Specifically, it will always be necessary to take account of the fact that one is

using data subject to error due to imputation by computing the correct estimates of statistics and the correct standard errors of these statistics using the equations we have presented above. Tempting shortcuts, such as arbitrarily picking one of the five imputed values, treating it as actual data, and incrementing the resulting standard errors by some percentage, are not legitimate; as we have seen, the extent of the error due to imputation varies across subsamples and across statistics.

This is but the first of a number of papers reporting on the Census multiple-imputation project. A companion paper evaluating the occupation imputation that is currently underway will be our next effort. It is fairly likely that the error due to imputation will be relatively larger for the occupation imputations than for the industry imputations because of the greater complexity of the occupational classification. But whether the reduction in precision will be substantial enough to seriously undercut the value of the imputed occupation data remains to be seen. As an aid to users of the imputed data, we also expect to produce a monograph that includes a number of "worked examples" in conjunction with release of a PUMS tape for 1970 containing imputed industry and occupation codes in the 1980 classification.

REFERENCES

Blau, Francine D., and Wallace E. Hendricks. 1979. "Occupational Segregation by Sex: Trends and Prospects." *Journal of Human Resources* 14:197–210.

Edwards, A. M. 1943. *Sixteenth Census of the United States: 1940. Population. Comparative Occupational Statistics for the United States, 1887 to 1940: A Comparison of the 1930 and 1940 Census Occupation and Industry Classifications and Statistics; a Comparable Series of Occupation Statistics, 1887 to 1930; and a Socio-Economic Grouping of the Labor Force, 1910 to 1940.* Washington, DC: U.S. Government Printing Office.

Kaplan, D. L., and M. C. Casey. 1958. *Occupational Trends in the United States 1900 to 1950.* Working Paper No. 5. Washington, DC: U.S. Bureau of the Census.

Pampel, Fred C., Kenneth C. Land, and Marcus Felson. 1977. "A Social Indicator Model of Changes in the Occupational Structure of the United States, 1947–1974." *American Sociological Review* 42:951–64.

Priebe, J. A. 1968. *Changes between the 1950 and 1960 Occupation and Industry Classifications—with Detailed Adjustments of 1950 Data to the 1960 Classifications.* Bureau of the Census Technical Paper No. 18. Washington, DC: U.S. Government Printing Office.

_____. 1985. "1970 Census Sample with Industry and Occupation Descriptions." Unpublished manuscript, U.S. Bureau of the Census.

Priebe, J. A., J. Heinkel, and S. Greene. 1972. *1970 Occupation and Industry Classification Systems in Terms of their 1960 Occupation and Industry Elements*. Washington, DC: U.S. Government Printing Office.

Rubin, Donald B. 1978. "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse. Pp. 20–34 in *Proceedings of the Survey Research Methods Section, American Statistical Association*. Washington, DC: American Statistical Association.

_____. 1987. *Multiple Imputation for Survey Nonresponse*. New York: Wiley.

Rubin, Donald B., and Nathaniel Schenker. 1986. "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81:366–74.

Rumberger, Russell. 1981. "Changing Skill Requirements of Jobs in the U.S. Economy." *Industrial and Labor Relations Review* 34:578–90.

Synder, David, Mark D. Hayward, and Paula M. Hudis. 1978. "The Location of Change in the Sexual Structure of Occupations, 1950–1970: Insights from Labor Market Segmentation Theory." *American Journal of Sociology* 84:706–17.

Subcommittee on Comparability of Occupation Measurement. 1983. *Alternative Methods for Effecting the Comparability of Occupation Measurement over Time: Report to the SSRC Advisory and Planning Committee on Social Indicators and the U.S. Bureau of the Census*. Washington, DC: Social Science Research Council.

Treiman, D. J., and K. Terrell. 1975. "Women, Work, and Wages—Trends in the Female Occupational Structure Since 1940." Pp. 157–99 in *Social Indicator Models*, edited by K. C. Land and S. Spilerman. New York: Russell Sage.

U.S. Bureau of the Census. 1904. *Twelfth Census of the United States 1900. Special Reports. Occupations at the Twelfth Census*. Washington, DC: U.S. Government Printing Office.

_____. 1973. *Census of the Population: 1970. Subject Reports. Final Report PC(2)-7A. Occupational Characteristics*. Washington, DC: U.S. Government Printing Office.

_____. 1976. *1970 Census of Population and Housing: Procedural History. Report PHC(R)1*. Washington, DC: U.S. Government Printing Office.

_____. 1981. *Census of Population Alphabetical Index of Industries and Occupations*. 2d ed. Washington, DC: U.S. Government Printing Office.

Vines, Paula, and John A. Priebe. 1988. *The Relationship between the 1970 and 1980 Industry and Occupation Classification Systems*. Washington, DC: U.S. Government Printing Office.

Williams, Gregory. 1979. "Trends in Occupational Differentiation by Sex." *Demography* 16:73–88.