ANALYZING THE DISCLOSURE REVIEW PROCEDURES FOR THE 1995 SURVEY OF CONSUMER FINANCES[1]

Gerhard Fries, Federal Reserve Board; Barry Johnson, Internal Revenue Service; and R. Louise Woodburn, Ernst and Young

Gerhard Fries, FRB, Mail Stop 153, Washington, DC 20551; m1gxf00@frb.gov

Key Words: Confidentiality, Imputation

A principal concern among survey practitioners is protecting the confidentiality of the survey respondent. This is important, not only for the direct purpose of keeping an individual's data anonymous, but also for the more global perception that it is 'safe' to participate in surveys. On the other hand, it is important to provide as much useful data as possible to policy makers and researchers. Adjustments made to the data in order to protect a respondent's identity could easily compromise the usefulness of the data. Thus, it is necessary to take measures to keep the integrity of the data intact. This paper is based on our experiences with the Federal Reserve Board's Survey of Consumer Finances (SCF), a triennial household survey that includes data on finances, employment and demographics. In this paper we analyze the disclosure procedures used in preparing the SCF data for public release. We focus on the actual procedures used in preparing the 1995 data for public release and investigate other procedures as well. Including this introduction, there are five sections. In the next section, we provide a brief summary of the SCF, covering the sample design, data collected, and disclosure issues. In the third section, we detail the disclosure strategy currently used in the SCF. An investigation to detect potential disclosure adjustments effects is presented next. We summarize our results and discuss their implications for future surveys in the last section.

Background on the SCF

The SCF is a triennial household survey sponsored by the Federal Reserve Board with cooperation from the Statistics of Income Division (SOI) of the Internal Revenue Service. Data are collected on household finances, income, assets, debts, employment, demographics, and businesses. The interview for the 1995 SCF was conducted via CAPI and averages about 90 minutes, but interviews of households with more complicated finances sometimes last several hours. An important objective of the SCF is to collect representative data to measure the distribution of household wealth in the U.S. In order to accomplish this, the sample is selected from a dual frame that is composed of an area- probability (AP) frame and a list frame (see Kennickell and McManus [1993] for details on the strengths and limitations of the sample design). The list frame is based on administrative records maintained by SOI. The list sample is stratified on an estimated wealth index, with higher indexes selected at a higher sampling rate.

Due to the sensitive nature of the financial questions, both unit and item nonresponse are concerns in the SCF. Frame data are available for the list sample that can be used for nonresponse adjustments in weighting (Kennickell and Woodburn [1997]). However, for the AP sample, only limited geographic data are available for this purpose.

To account for item nonresponse, missing values are multiply-imputed (Kennickell [1992] and Rubin [1987]). For the 1995 SCF, the respondent has many options for answering a given question, he can: 1) give a specific value, 2) decline to answer ("refuse" or "don't know"), or 3) choose a range, either from a range card provided by the interviewer or a self-constructed range. When a "don't know" or "refuse" reply is given, the CAPI program iteratively tries to narrow in on a range (see Kennickell [1996]). The imputations for the range responses are constrained by the range interval boundaries. The imputation approach involves iteratively estimating a sequence of large regression models to draw values for the missing values based on variables that are available for a given respondent. The result is an imputed data set that preserves the distributions and relations found in the non-imputed data. For all of the survey variables, a shadow variable is included that indicates the status of the original data, e.g., range response, "don't know", "refuse", etc. The imputation machinery is used in the disclosure avoidance preparation of the public use file as described below.

Means to estimate both sampling and imputation error are also included in the survey database. Estimates of the variance due to imputation are computed using five imputation replicates ("implicates"). Estimates of the variance due to sampling are computed using the bootstrap method with 999 bootstrap replicates. (A thorough reference for the bootstrap method is Shao and Tu [1995]).

Disclosure Issues in the 1995 SCF

This section details the data review and subsequent actions taken to reduce the possibility that the identity of an SCF respondent can be determined using the publicly released microdata. Protecting the privacy of survey participants has become an increasingly challenging responsibility as the availability of both personal data in the market place and computer technology continue to expand rapidly (Felligi, [1997]). The use of administrative files as a part of the SCF brings an added legal responsibility for protecting the identity of those included in the list portion of the sample; because they are sampled from a known, finite population, there is also an added risk of re-identification.

While no direct identifiers (name, address, SSN) are collected as a part of the survey, there are other potentially identifying variables, such as occupation, sex, age, primary sampling unit, and marital status. These variables, when combined with the detailed financial and household information collected might be used to identify individual respondents. This is especially true of individuals whose combined demographic and financial characteristics make them relatively rare in the general population (de Waal and Willenborg, [1996]). Recent advances in computer technology and sophisticated record linkage software, especially advances in probabilistic record linkage (Scheuren and Winkler, [1997]) only serve to increase the possibility that the identity of a survey respondent could be discovered if adequate steps are not taken to limit disclosure.

Disclosure Adjustments

There are many techniques that have been used to minimize disclosure for public use microdata files. The priority of these techniques has been to protect the identity of individual respondents, while at the same time preserving the integrity and usefulness of the original data (Fries and Woodburn, [1994]). Potential masking procedures include top/bottom coding, data swapping, blurring, adding random noise and blank and impute (OMB, [1994]). Another potential method is to create synthetic data using model based imputation techniques which utilize the original data as the underlying model (Rubin, [1993]). Interest in use of artificial data generated from true data has gained in popularity in recent years. Researchers at the Census Bureau, and elsewhere are working to develop this technique (Evans, Moore, and Zayatz, [1996]). This approach has also been explored using data from the 1995 SCF (Kennickell, [1997]).

While the goal of the public release file was to release as much data as possible, inevitably, some data items were not included in the final data set. These included certain details about a respondent or spouse's marital history and components of the sample design and weight components. In general, variables which were suppressed were not directly related to the main purpose of the survey. In a few cases, new variables were created using survey responses to enable users to conduct relevant research without releasing actual reported values which might provide direct clues to a respondent's identity.

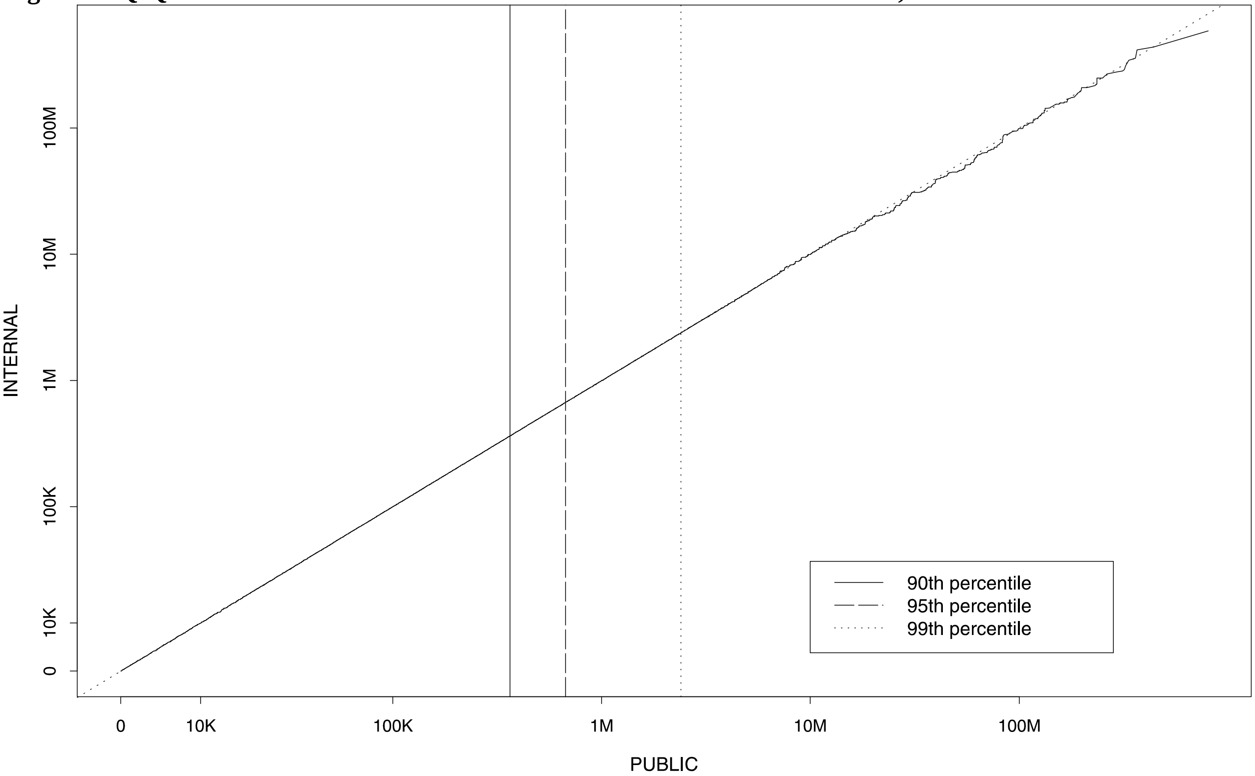
Initial review of the data involved graphical analysis of monetary data items. The use of scatter plots showing the variable of interest versus an indication of the sampling strata were particularly useful for identifying responses which were 'unique', both overall and for selected subgroups (Fries and Woodburn, [1995]). Frequency tables for each of the ordinal and discrete variables were also used extensively to determine the number of responses in each category and to estimate the disclosure potential of each response. In addition, a team member, assuming the role of an intruder, was employed to evaluate the identifying potential of a group of variables.

Monetary data items and discrete variables were treated using a combination of disclosure techniques. In earlier releases of SCF data, blanking out original values of monetary data items which were identified as sensitive and replacing them with imputed values played a key role in the overall disclosure review procedure. This was implemented for selected variables by first identifying extreme values and then determining whether or not those values posed a significant disclosure threat. For the 1995 study, this method was largely abandoned in favor of an approach based on the emerging ideas about synthetic data. More than 350 cases were selected, both purposefully, based on specific characteristics, and randomly. All of the monetary data items for these cases were then blanked out and fully imputed using the FRITZ software developed by Arthur Kennickell (Kennickell, [1997]). The imputations for responses that were originally given were constrained to fall within a predetermined range of the original value, for example +/- 11 percent, much like adding random noise, except that by using the imputation software, relationships between financial values were preserved. All variables that were previously imputed were reimputed. Finally, monetary data items for all cases were rounded and large negative values were bounded at -$1,000,000.

Likewise, many different strategies were used to reduce the sparseness of the responses to the non-continuous variables. Top and bottom coding were used extensively for ordinal variables describing years, numbers of items owned, frequencies of events, etc. Rounding was also used to further mask certain dates. Categories were collapsed for many discrete variables so as to limit the detail on the final file. In general, categories containing less than 3 responses were combined with other related categories; variables with a strong potential for re-identification, such as occupations, were collapsed even further.

Data swapping also played an important role in the overall disclosure prevention strategy of the SCF. It is an attractive tool because it is easy to implement and can be

Figure 1. Q-Q Plot of Public Dataset Net Worth vs. Internal Dataset Net Worth, 1995 SCF



used on sensitive variables without disturbing others (Moore, 1996). For the public release of the 1995 SCF, data swapping was used on the geography variables (4-level Census region and 9-division Census area). These values were swapped with records containing similar characteristics on key variables in an attempt to preserve most univariate statistics for the overall data set as well as for important subsets of the population.

Analysis of Disclosure Adjustments

This section will review different analyses that were conducted on the public data to detect potentially important disclosure adjustment effects. A high priority in the design of the adjustments is to avoid changing the underlying integrity and usefulness of the data. This discussion will concentrate on the measure of wealth (net worth), although assets and debts were reviewed as well. The analyses performed include overall distributional comparisons and mean comparisons for different demographic groups by Census region and division.

Figure 1 shows a Q-Q plot of net worth estimated from the internal data versus net worth from the public data. The inverse hyperbolic sine transformation with a scale parameter of .0001 (see Burbidge, Magee, and Robb [1988]) was used. This transformation eliminates exaggerations near zero and compresses large spreads in the tails of the distribution. To avoid disturbing aberrations caused by the very few values of negative net worth, all negative values were set to zero. Also shown are lines corresponding to the 90th, 95th, and 99th percentiles of the net worth distributions. A Q-Q plot lying on the 45 degree line would indicate that the distributions are identical.

An investigation of the figure reveals very small distortions in the upper tail for the top one percent of the distribution. Most of these deviations fall below the 45 degree line indicating perhaps a small relative increase in the estimated net worth of the group from the public data compared to that from the internal data. The deviation at the very top arises from the difference in maximum values for both distributions. The rest of the plot conforms closely to the 45 degree line.

Table 1 shows estimates for aggregate holdings and percent of the total for net worth for both the public data and the internal data by different percentiles. Also included are standard errors with respect to imputation and sampling.[2] Estimates and standard errors are very similar throughout the table. Consistent with results from the Q-Q plot, there is a slight increase (95.5 Billion) for aggregate holdings, as well as for percent of total (0.4), when comparing results from the households between the 99.5 and 100 percentiles for the public data versus the internal data. However, these differences are not significant. A similar, but smaller increase (7.4 Billion) exists for the aggregate of the 99.0-99.5 percentile group. Again, this difference is small compared to the standard error for the estimate. In order to review the effects of data swapping for the geography variables, point estimates by Census region and division for means and medians of net worth, debt, and assets were reviewed by income, education status, and age

Table 1. Proportion of Total Net Worth Held by Different Percentile Groups: 1995 SCF, Internal and Public Use Datasets. All dollars values given in billions of 1995 dollars.

	Percentiles of the net worth distribution									
	All Families		*0 to 89.9*		*90 to 99*		*99 to 99.5*		*99.5 to 100*	
Networth	$	% of total	$	% of total	$	% of total	$	% of total	$	% of total
Internal	20,519.8	100.0	6,472.8	31.5	6,821.0	33.2	1,566.0	7.6	5,650.0	27.5
	1,398.3	*0.0*	*317.9*	*1.8*	*529.5*	*1.4*	*250.7*	*0.7*	*688.6*	*2.0*
Public	20,629.3	100.0	6,478.8	31.4	6,821.6	33.1	1,573.4	7.6	5,745.5	27.9
	1,419.3	*0.0*	*316.3*	*1.8*	*532.2*	*1.4*	*248.4*	*0.7*	*718.4*	*2.1*

Standard errors due to imputation and sampling are given in italics.

groups. The estimates for the medians, as expected, are affected very minimally. The largest differences occur in mean net worth for families with income greater than $125,000. Analysis reveals that even these largest differences are, in fact, relatively "small" and insignificant.

Table 2 contains means for net worth calculated from the public data for income categories by 9-division Census area. Figure 2 is a graph of the data points in these cells plotted versus the corresponding data points from the internal data. A 45 degree line is added as a viewing tool to show how "different" the points are from each other. The labeled points are all points from the greater than $125,000 income category.

Table 3 includes the point estimates and associated standard errors for one of these cases — the Pacific Division. The difference (121.2 Thousand) is not significant since the standard errors (almost 400 Thousand) for each individual estimate are in themselves on the order of three times larger than this difference. The results are similar for the other highlighted points in Figure 2. It is comforting that for such small disaggregated sub-populations of the total U.S. population, our results indicate that there are no obvious disclosure effects relating to the swapping adjustments applied to create the 1995 public data.

Table 2. Mean Net Worth by Income Category and Census Division:
Public Use Dataset. All dollar values given in thousands of 1995 dollars.

Income Category	New England	Middle Atlantic	South Atlantic	East S. Central	West S. Central	East N. Central	West N. Central	Mountain Division	Pacific Division
< 35	94.8	84.1	55.9	52.6	45.0	74.0	80.5	72.9	100.1
35-50	142.7	157.3	137.0	106.7	89.6	119.9	114.5	167.8	176.0
50-75	273.6	281.0	208.8	208.0	166.6	176.3	182.8	191.0	219.4
75-125	334.3	349.3	520.4	333.7	420.0	310.0	436.0	540.6	422.0
> 125	2016.8	1836.1	2101.8	2475.2	1354.3	1583.0	2571.7	1965.8	2295.5

Figure 2. Mean Net Worth by Income Categories and Region, 1995 SCF

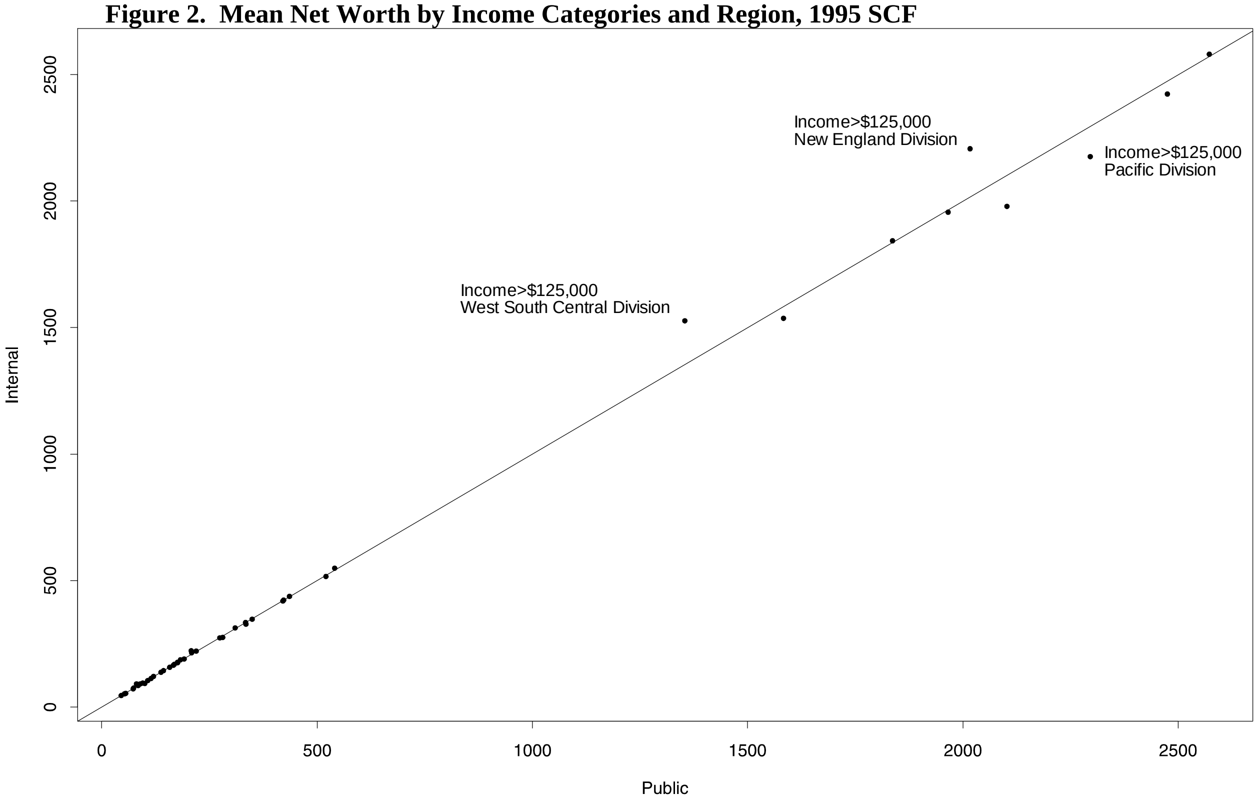


Table 3. Net Worth Estimates for the Pacific Division - Income > 125,000:
1995 Public Use SCF. All dollar values given in thousands of 1995 dollars.

Dataset	*Mean Estimate*	*Standard Error*
Internal	2,183.7	371.1
Public	2,304.9	394.9

Table 4. Percent Median Ratio of Debt Payments to Family Income for 1989, 1992, and 1995 SCF'S in Percent for Families with Income < 10,000.

Dataset	*1989*	*1992*	*1995*
Internal	22.0	13.2	16.0
Public	22.9	13.6	16.7

The evidence so far is encouraging, but one can hardly expect that any study performed on both the public data and the internal data will always yield identical results. This is especially so if one considers the thousands of variables in the public dataset to which users can apply an endless variety of analyses. The following recent example is likely not unique.

As a starting point, analysts using the public dataset will often attempt to reproduce results which have appeared in Federal Reserve publications (e.g. *Federal Reserve Bulletin* articles). In general, it appears that most such users who relate their experiences to our staff are satisfied and are able to arrive at estimates or conclusions similar to our own. An outside user was concerned with discrepancies between her calculations and those reported in an article published in the January 1997 *Federal Reserve Bulletin*. Table 4 shows the median ratio of debt payments to family income for families with less than $10,000 of income, estimated from the 1989, 1992, and 1995 SCF's using both public and internal data. The question of concern was whether these differences in the median are reasonable. After careful review, it was determined that the distribution is relatively "thin" around the median and that rounding used in the disclosure adjustments can easily account for these differences.

Conclusions and Future Plans

This paper provides two encouraging results. First, the use of controlled imputation of all monetary values for an important subgroup of respondents seems to preserve important relationships and characteristics of the original data. Second, for the analyses presented here, there were no significant differences between results produced using the internal data and those produced using the public data.

Future research should extend this analysis to cover the effects of disclosure adjustments on more types of analyses, including econometric modeling. Further investigation into the effectiveness of these adjustments in protecting respondent identities is also warranted, given growing concerns over individual privacy. Finally, it would be useful to compare the effectiveness of the disclosure techniques described in this paper with those applied to earlier SCF's.

Acknowledgements

The views presented in this paper are those of the authors alone and do not necessarily reflect the views of the Board of Governors of the Federal Reserve System, the Internal Revenue Service, or Ernst and Young. The authors would like to express their gratitude to all of the Survey of Consumer Finances staff for their support with the disclosure review implementation, and especially Kevin Moore and Amy Stubbendick for outstanding research assistance. A special thanks to Arthur Kennickell for invaluable guidance and comments.

Endnotes

1. The full version of the paper will be available on the FRB SCF web site: **www.bog.frb.fed.us/pubs/oss/oss2/scfindex.html**

2. The standard error for statistic X is estimated as $SX_{tot} = \{(6/5)*SX^2_{imp} + SX^2_{samp}\}^{1/2}$, where the imputation variance SX^2_{imp} is given by $SX^2_{imp} = (1/4) * \Sigma_{i=1 \text{ to } 5} (X_i\text{-mean}(X))^2$ and the sampling variance SX^2_{samp} is given by $SX^2_{samp} = (1/999) * \Sigma_{r=1 \text{ to } 999}(X_r\text{-mean}(X))^2$.
For the imputation variance, the mean function is taken with respect to all five implicates. Since we have computed bootstrap weights only for the first implicate, for the sampling variance calculations, the mean function is taken with respect to the 999 bootstrap replicates of the first implicate.

References

Burbidge, J.B., L. Magee, and A.L. Robb [1988] "Alternative Transformations to Handle Extreme Values of the Dependent Variable," *Journal of the American Statistical Association,* Vol. 83, No. 401, pp.123-127.

Evans, T., R. Moore, and L. Zayatz [1996] “New Directions in Disclosure Limitation at the Census Bureau,” Unpublished Manuscript.

Felligi, I. [1997] address given at 1997 Record Linkage Techniques, Washington, DC.

Fries, G., and R.L. Woodburn [1994] “The Challenges of Preparing Sensitive Data for Public Release,” *Proceedings of the Section on Survey Research Methods*, 1994 Annual Meeting of the American Statistical Association, Toronto, Canada.

Fries, G., and R.L. Woodburn [1995] “Using Graphical Analyses to Improve all Aspects of the Survey of Consumer Finances,” *Proceedings on the Section of Survey Research Methods*, 1995 Annual Meeting of the American Statistical Association, Orlando, FL.

Jabine, T.B. [1993] "Statistical Disclosure Limitation Practices of United States Statistical Agencies," *Journal of Official Statistics*, Vol. 9, No. 2, pp. 427-454.

Journal of Official Statistics [1993], Confidentiality and Data Access, Vol. 9, No. 2.

Kennickell, A.B. [1997] "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances," paper presented at 1997 Record Linkage Techniques, Washington, DC.

Kennickell, A.B. [1991] "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," *Proceedings of the Section of Survey Research Methods*, 1991 Annual Meeting of the American Statistical Association, Atlanta, GA.

Kennickell, A.B., and D.A. McManus [1993] "Sampling for Household Financial Characteristics Using Frame Information on Past Income," *Proceedings of the Section of Survey Research Methods*, 1993 Annual Meeting of the American Statistical Association, San Francisco, CA.

Kennickell A.B., D.A. McManus, and R.L. Woodburn [1996] "Weighting Design for the 1992 Survey of Consumer Finances," Federal Reserve Board Working Paper.

Moore, R.A. [1996] “Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets,” Bureau of the Census Statistical Research Report Series No. RR96/04.

Office of Management and Budget [1994] “Report on Statistical Disclosure Limitation Methodology,” Statistical Policy Working Paper 22.

Rubin, D.B. [1993] "Discussion: Statistical Disclosure Limitation," *Journal of Official Statistics*, Vol. 9, No. 2, pp. 461-468.

Rubin, D.B. [1987] *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons, Inc. .

Winkler, W., and F. Scheuren [1997] "Analysis Issues in the Presence of Linkage Errors," paper presented at 1997 Record Linkage Techniques, Washington, DC.

de Waal, A.G., and L.C.R.J. Willenborg [1996] “A View on Statistical Disclosure Control for Microdata,” *Survey Methodology*, Vol. 22, No. 1, pp 95-103.

Wilson, O., and W.J. Smith Jr. [1983] "Access to Tax Records for Statistical Purposes," *Proceedings of the Section of Survey Research Methods*, 1983 Annual Meeting of the American Statistical Association, Toronto, Canada.