Data Comparability Among Federal Household Surveys

Thesia I. Garner, Jeanne M. Hogarth, Richard D. Miller,
William Passero, Nell Sedransk

Introduction

The U.S. federal government collects a wide array of household survey data including the Consumer Expenditure Survey (CEX), the Survey of Consumer Finances (SCF), the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), the Annual Housing Survey (AHS), the National Food Consumption Survey (NFCS), and the National Medical Expenditure Survey (NMES). All are used to collect data on the economic well-being of individuals, economic units, families, and households. However, differences exist in sampling designs, timing of the surveys, questions asked, and the types of data collected. These differences make it difficult for researchers to use the richness of these federal household surveys as a combined set. Previous researchers (Kilss and Scheuren 1978; Wolfson et al. 1989) have produced combined data sets from separate national data files using exact record and statistical matches. Continuing in this tradition, the U.S. Department of Labor's Bureau of Labor Statistics (BLS), within the Divisions of Price and Index Number Research (DPINR) and Consumer Expenditure Surveys, embarked on a long-term project in 1988, with the cooperation of other federal agencies, to produce integrated multiple federal household survey data sets. A first step in this project is to identify household surveys with comparable fields for similar populations and to examine the data in each of the files. The primary aim at the integration stage is to follow a more statistically principled approach to the integration of data, such as those proposed by Rubin (1986) and Rosenbaum (1989), than has been followed in the past (see Scheuren 1989).

An integrated or augmented data set would serve

Thesia I. Garner is a research economist with the Division of Price and Index Number Research (DPINR) in the Bureau of Labor Statistics (BLS). Her work focuses on economic well-being and health economics. She received her Ph.D. from the University of Maryland in consumer economics.

Jeanne M. Hogarth is an associate professor in Department of Consumer Economics and Housing at Cornell University. She received her Ph.D. in family and consumer economics from the Ohio State University. Her research interests focus on issues related to retirement decision making.

Richard D. Miller is an econonmist with the DPINR in the BLS. His arch interests focus on health economics. He received his MA in economics from the University of Wisconsin

William Passero is a graduate of Brown University and has been an economist with the Division of Consumer Expenditure Surveys at BLS for the last 15 years. During that time he has been involved in all aspects of the Consumer Expenditure Survey, including design, processing, data dissemination, and analysis.

Nell Sedransk is a 1989-90 National Science Foundation American Statistical Association Fellow at the BLS. Currently, she is on leave from the University of Iowa. Her research focuses on mixture distributions, experimental design, and Bayesian design and inference. She received her Ph.D. in statistics from Iowa State University.

several purposes within the federal household survey community. Sponsoring agencies are concerned that the data collected in their surveys are both accurate and genuine. Much time and effort is spent testing and verifying data prior to public release. In the CEX program, one of the validation procedures involves comparing estimates derived from the CEX with estimates from other sources that collect similar data. Because of conceptual and methodological differences, these comparisons can be crude and inexact In creating an integrated data set of the sort described above, issues of data comparability can be dealt with more formally and systematically, allowing refinement of the validation procedure when dealing with other overlapping surveys.

A second purpose served by such an augmented data base is to facilitate research by incorporating information on household units not available from any one survey. Thus, researchers should be better able to address various questions using the most appropriate data available, not being limited to only one source. For example, various studies have explored the economic well-being of individuals and families in terms of income patterns, asset management behaviors, and consumption and expenditure patterns, as well as their access to community and in kind resources. These measures of economic wellbeing are inextricably linked to each other. Unfortunately, because of problems in attaining comparable data, few researchers have been able to explore a broader picture of the interrelationships among income, assets, community resources, and expenditures. An augmented data set can be used to assess the economic status of households and to develop appropriate local, state, and federal policies, as well as to design consumer education programs.

The purpose of this paper is to review the activities within the BLS during the first stages of the data comparability project. In it we describe how a data base can be created from federally collected micro data sets, using the 1983 Consumer Expenditure Survey and the 1983 Survey of Consumer Finances, and note how such a data base can be used to study the economic well-being of individuals and families.

History of Project

For the American Council on Consumer Interests (ACCI) annual conference in 1988, Garner organized a session on "Mining the Richness of Federal Data Bases" that explored how consumer economics researchers could make use of data from the CEX, the SCF, and the SIPP. These surveys are administered by the BLS, the Federal Reserve Board, and the Bureau of the Census, respectively. The session stimulated discussion among members regarding how to combine various federal household surveys to supplement information in any one survey and to address issues not answerable from a single data base.

Further discussions among BLS, Federal Reserve Board, and Census staff after the conference resulted in a commitment to begin a project to develop a data base incorporating data from multiple federal household surveys. Staff at the BLS would coordinate and administer the project, and would produce data files in which comparable data fields from the surveys would be drawn

In the fall of 1988, Hogarth approached the Bureau concerning the possibility of joining the agency during her sabbatical, beginning in the summer of 1989. After discussions within the Bureau, it was decided, given Hogarth's interesting world with multiple data sets to study retirement issues, that her addition to the household data comparability/data matching project would be valuable. It was at this time that the SCF was selected as the first survey to be matched with the CEX One of the major interests within the Bureau and among other users of the CEX data has been the quality of the CEX income, asset, and liability data. The SCF is specially designed to obtain information on these financial variables. Hogarth's primary responsibilities would be to identify the comparable variables, to design comparable data sets from the two surveys, and to begin the process of data matching, time permitting. Hogarth joined the Bureau for five months in July 1989.

BLS staff members, including Passero and Garner, had been working on the project for several months prior to Hogarth's arrival. These individuals prepared a survey matrix with descriptions of each survey's administration, design, and data (the Survey Matrix is discussed later). Sedransk, a statistician, joined the project in the spring of 1990 to provide assistance on issues related to data matching. Miller created the comparable data files upon which the initial match would be based.

Related Literature

Various researchers have tackled the issue of comparing data from one survey to another. However, most of the comparisons have been descriptive in nature, comparing results from one survey with those of another for "base-line" purposes; the data sets are not combined in any way. Other researchers have focused their attention on creating comparable data files that can be pooled in order to facilitate research, for example, on cross-national comparisons. A third type of comparable data files results from matching or merging data from two or more sources. Selected studies from each of these types of comparable data sets are briefly reviewed. As a final part of this section, a review of several articles in which data match/merging issues have been addressed is presented.

Compared Data

Numerous studies have compared results from one survey to those of another in order to provide some basic standard of comparison or validation. Carlson and Dalrymple (1986) and Ku and Dalrymple (1987) worked with data from SIPP, the Food and Nutrition Service (FNS), and the Panel Study of Income Dynamics (PSID). Comparability issues they encountered included differences in accounting periods, the definition of the sampling unit, geographic coverage, and skip patterns.

Avery and Kennickell (1989) and Curtin, Juster, and Morgan (1987) compared wealth estimates from the 1983 SCF with IRS, CPS, Flow of Funds, SIPP, and PSID data. Issues of weighting, item comparability, and item measurement were noted.

Buse (1986) and Buse, Cox, and Glaze (1986) used the 1977-78 National Food Consumption Survey (NFCS) and the 1972-73 CEX to determine the consistency in reported expenditures on food. They concluded that there was no significant difference in expenditures reported in the NFCS and the 1972-73 CEX. Buse (1988) has also examined a number of data sets in terms of their use for food demand estimation. He identified data sets available, and noted the conceptual issues and statistical problems involved when comparing results based on both time series and cross-sectional data sets.

Pooled Data

Data used to produce comparable files can be collected by using the same survey administered in different geographic regions or by using different surveys and organizing the comparable data into like fields. An example of the latter is the Luxembourg Income Study (LIS). The LIS was created specifically to allow researchers, with interests in the economic well-being of households, to use comparable data from several countries from a single data file (Smeeding and Rainwater 1988). The LIS data set contains CPS-type data from 10 countries. Many of the more than 40 income and 30 demographic variables are not exactly aligned. However, these data are an excellent first step at creating a broad data base that allows the research community to study such topics as the economic status of single parent families (Hauser and Fischer 1985), the economic status of the elderly (Achdut and Tamir 1985), public sector transfers and income taxes (Aguilar and Gustafsson 1987), and equivalence scales (Buhmann et al.1988).

Matched/Merged Data

The third type of comparable file results from matching records and merging data from two or more sources. Exact matches can be used to supplement information from a survey with administrative data from another source. Statistical matches result when household or individual data from two different sources are linked by common variables, but different households and individuals are represented in the data sets. Data from one survey can be used to provide more complete information for certain fields in another survey. Or, information not available in one survey can be supplemented with data from another survey. Exact matching and statistical matching can be used in conjunction to produce an enriched final data file. In this section several of the matched/merged data set projects are reviewed.

Early work on the matching/merging procedures was carried out by Budd (1972), Alter (1974), and Okner (1982). This early work relied on creating a matrix of characteristics and matching observations from two surveys within the cells of this matrix. One of the critical issues addressed was that of the cell weights. Also, there were critical comments on assumptions of the joint distribution given two marginal distributions and the issue of unique versus multiple matches. Okner's reply to these and other comments was that the criteria for judging a matching and merging procedure is how well the resulting data fit the research need for which they are intended. Thus, different matching criteria may need to be used depending on the question under study.

More recent efforts to match U.S. data include the 1985 CPS-Statistics of Income--CEX Statistically Matched Data Files (Lewin 1989). The data files were created for the U.S. Department of Treasury by Lewin, a consulting firm in Washington, D.C., matching data from the three sources using a 5,760 cell characteristics matrix.

The Census Bureau, the Social Security Administration, and the Internal Revenue Service (IRS) have matched the March 1973 CPS with data from Social Security benefit and earnings records and from federal income tax records (Kilss and Scheuren 1978; Scheuren 1980). Exact matching by Social Security number was used (this is not an option to the general research community). This comparable data file was used by Greenlees, Reece, and Zieschang (1982) to impute values for wages and salaries based on federal income tax returns to non-respondents in the CPS data.

Wolfson et al. (1989, p.36), researchers with Statistics Canada, constructed the Social Policy Simulation Database (SPSD) with "realistic, albeit synthetic" data on individuals in households for use with their Social Policy Simulation Model (SPSM). The SPSD was constructed from four major sources of microdata: the 1984 Canadian SCF, personal income tax returns, unemployment insurance claim histories, and the Family Expenditure Survey. Techniques used to construct the SPSD included controlled blurring, integrated weighting, categorical matching conversion, micro-record aggregation, and stochastic imputation (see Wolfson et al. 1989 for a discussion of these procedures). With the publicly available SPSD, the researchers state that anyone can perform microsimulation analyses of tax and transfer programs in Canada on a personal computer. They note that the sophistication of this data base system approaches or exceeds that of current models used by federal departments in Canada.

Scheuren (1989) evaluated the SPSD design and commended Wolfson and his colleagues for achieving a breakthrough by making these data and their model accessible to researchers at far lower costs than had been previously possible. However, he expressed concern over the imputations made and statistical matching techniques used. Scheuren suggested that rather than blending various techniques, it would be better to follow a more statistically directed approach to the integration of data (Rubin 1986 and 1987; Paass 1989).

Data Matching Issues

Discussions of techniques and issues surrounding matching or merging two or more data sets have been going on for the past 20 to 30 years: Issues addressed include the careful alignment of variables (Ruggles and Ruggles 1974), the use of consistency scoring and distance functions to evaluate the quality of a match (Ruggles and Ruggles 1974; Ruggles, Ruggles, and Wolff 1977), and continued development and testing of techniques to improve the matches made. Discussions of this latter issue has been led most recently by Scheuren (1985), Rubin (1986) and Rosenbaum (1989).

Ruggles and Ruggles (1974) addressed the need for careful alignment of the variables common to the candidate data sets. They appear to have been the first to suggest the use of a distance function for matching, and suggested use of regression or consistency scores to rank observations within the cells to facilitate the matches. Distance functions could also be used to match an observation at the boundaries of the cell with other observations near the boundary in an adjacent cell to produce a better match.

Ruggles, Ruggles, and Wolff (1977) further explored some of the goodness of match issues by testing a matching strategy using the 5% and 15% Public Use of the match (which data set is primary versus secondary) of the match (which data set is primary versus secondary) versus a "symmetrical" match can matter. Additional considerations they noted include aligning the unit of analysis (individuals versus households verses families), selection of matching variables and allowance for matching within a range of these variables, and hierarchy of sorting (which criteria are mandatory and which can be relaxed in order to make a match). They concluded that a sorting and merging procedure "can provide reliable synthetic data sources for many kinds of statistical applications" (Ruggles et al. 1977, p. 428).

Scheuren (1985,1989) has long been a supporter of linked multiple data bases and has systematically noted the methodological issues to consider in the design and creation of such files. Under the rubric of structural design considerations he has emphasized that agency and user support for the data matching project need to be very strong and continuing. Technical or matching design considerations focus primarily on personnel, computer software, and data. Scheuren has argued that it is imperative that statisticians get involved in the matching process very early to insure that the donating data sets are organized for optimal matching, that resulting matched files maintain statistical properties

similar to the donor data sets, and that the errors in the data base are examined and documented. He has identified several data issues which should be addressed during the planning process before the actual matching begins: identification of the linking variables to be used, preprocessing of the data, determination of the matching rules, and how to handle indeterminate links (e.g., non-links, multiple links, potential links). After matching, it is mandatory that a sensitivity analysis of the results to the assumptions in the linkage process be made.

Rubin (1986) has recently directed his attention to the development of statistical matching techniques which allow users to test the sensitivity of inference to assumptions being made when creating a matched file. His approach to file matching is known as file concatenation with adjusted weights and multiple imputations. Use of multiple imputations enables the user to assess the effect of uncertainty in the match through the examination of the variation in answers as each set of imputed values is treated as "real."

The concern over goodness of the match has continued and additional techniques have been explored to improve the matches made. Rosenbaum (1989) draws on network theory (finding a flow of minimum cost in a network) to obtain optimal (versus "greedy") matches. He suggests using propensity/consistency scores as distance functions, but takes the technique a step further by minimizing the total distances between matched pairs. In "greedy" matching, matches are sequential and the first match made can remove both observations from the process. Optimal matching is an iterative process in which matches may be broken up and re-matched in order to minimize the total distances between all observations.

Techniques for Creating Composite Data Bases

Prior to discussing merging techniques, some notes on notation will be helpful. For simplicity, the case of merging two surveys is considered. For Survey 1, let the sampled units be indexed by i=1,...,I; for Survey 2, let the sampled units be indexed by j=l,...,J; note that there are no observations common to both surveys. Accordingly, let x_i or x_i be the vector of values for variables common to the two surveys, such as demographic data; let y_i be the vector of observed values for items unique to Survey 1; and let z. be the vector of observed values unique to Survey 2. The objective is to create a large file of I+J "pseudo" units of the forms: $(y,x,z^*)_i$ and $(y^*,x,z)_j$, where the values for z* and y* are imputed or taken from paired observations from the other survey. If each of these surveys uses a complex sample design and makes adjustments for refusals to participate, etc., each observation may have an associated weight, w_i or $v_{i.}$

Statistical Matching

In statistical matching, the common x variables are used to create a matrix where observations in the same cells are matched with each other using a variety of matching strategies. For statistical matching to be successful, it is necessary to select households with characteristics similar enough so that the merged information is consistent with all the known information in both files. Statistical matching is valid only in "dense" data sets, with many observations.

The basis for most matching procedures is the selection of one or more "nearest neighbors" from the second population for each unit in the first population, where "nearest" refers to some distance measure using some or all of the elements of x. Although these methods are generally "model-free," the choice of covariates and the weights assigned to each in calculating distance are critical in choosing the particular matches and in determining the quality of the matches. Often, matching procedures are used in an asymmetrical situation where the "target" population is modest in size compared with the "control" population. One of the advantages of statistical matching procedures is the opportunity to generate multiple matches to allow a direct estimate of variance which is not model-dependent.

A major concern in statistical matching is maintaining the variability in the linked data set. Suggestions to control for artificially shrinking variances and regression toward means, especially inappropriate means, include using multiple imputations or matches in the data that can then be used to assess the uncertainty and maintain the variance in the system.

Rubin (1986) suggested using adjusted weights and multiple imputations based on differing assumptions to estimate the "missing values" of the y and z variables in a symmetric match. He provided an example in which the partial correlations between y and z were set at two different levels and regression estimates were generated and matched to observed variables. He concluded that the assumptions about the partials can lead to very different imputed values

and thus very different results and conclusions. Rosenbaum (1987) also suggested using multiple imputation techniques to generate a sensitivity analysis on matched data. Again, different assumptions about the unobserved covariates may cause substantially different results.

Model-Based Imputation

To some extent, the "blank" observations of the z's in Survey 1, or the y's in Survey 2, are "missing data" and can be imputed using a variety of imputation techniques. Most of the model-based imputation methods assume a sufficiently large subpopulation with complete information, i.e., of the form (y, x, z), so that units with partial information do not require much extrapolation. This approach depends upon the degree of dependence of the observations to be imputed on the available covariates, upon the adequacy of the particular model chosen, and upon the sufficiency of the range of values of covariates used in fitting the model. Estimates of variance are usually model-based as well.

The strengths of the model-based imputation methods rest in the stability the model gives to the predicted or imputed values, at least to the extent that the model is appropriate, and in the basis for calculation of variance estimates under the model assumptions. The principal weakness when used for data merging is that in the absence of complete observations (y,x,z), no direct validation is possible, nor can the variance estimates be checked. There is also the obvious limitation that covariates for the prediction equation can only be drawn from x.

Composite Data Base Creation: Matching the CEX and SCF Data Bases

The matching of data from two or more surveys involves several steps. First is the identification of similarities and differences in survey design including sampling and questionnaire design. Second is the identification of variables to be used in the matching process. Third is the creation of aligned data files for each survey. Fourth is the matching process itself. And the fifth step is the evaluation of the matches in terms of their quality and sensitivity to assumptions made during the matching process. For the CEXSCF match, steps one through three have been completed. Step four is in progress. Steps four and five will have been completed by winter 1990-91. In this section of the paper, we review the first steps in, the matching of the CEX and SCF data bases.

Similarities and **Differences in the Surveys**

Our first step in aligning the CEX and SCF data sets involved developing a matrix that allowed us to comparevarious aspects of the surveys and to subsequently assess their comparability. The survey matrix is available from the authors. The matrix is divided into four sections with descriptions of the following. (1) the surveys in general, sampling issues, respondents, sources of error, and weighting; (2) survey variables; (3) published tabulations; and (4) public use microdata.

The most recent SCF data are for 1983 and 1986;1989 data have been collected but are not yet available. The CEX has been ongoing since 1980; the most recent available public use data are for 1988. Although the primary objectives of the two surveys are different, both collect similar financial and demographic information on households in the U.S. and both used interview instruments to collect the data. (The diary survey component of the CEX is not included in this part of the data comparability project) The population coverage is slightly different; both surveys collect data from the U.S. civilian non-institutional population, but the CEX also collects data from a portion of the institutional population. The sampling frames for both the 1983 SCF and CEX surveys were generated from the 1970 Census 100 percent detailed file, using Primary Sampling Units (PSUs); however, the CEX sampling frame was augmented for new construction. A special stratified high income sample based on IRS files was included for the SCF. The CDC interview survey used a panel rotation design, targeted at approximately 5,000 consumer units quarterly. The 1983 SCF sampled 6,062 in the area probability sample and 459 in the high income sample. Thus, although there are differences in the sampling schemes, the SCF area probability sample was closely enough aligned to the CEX sample that the data could be matched with some degree of integrity.

In 1986, the other comparable year for the SCF and CEX, the sampling frame for the CEX was based upon the 1980 Census 100% detailed file. For the CEX, rural areas within supplemental PSUs were temporarily eliminated from October 1981 through December 1983; rural households were included in the 1986 sample. The 1986 SCF was a re-interview of the 1983 respondents and their spouses. The 1986 survey was a telephone survey and resulted in 2,822 interviews. The 1986 SCF

was more limited in scope than the 1983 survey.

Timing of the data collection was another key point for data alignment. The CEX uses a panel rotation design and has been on-going since 1980. A panel includes all consumer units who are interviewed in the same quarters over time. The survey is designed such that the consumer unit is visited five times over a 13month period, with the last four interviews occurring once per quarter, expenditure data collected in the first interview is used for bounding purposes only. The CEX is used to ask for expenditures over the past three months; in the case of income, the second and fifth CEX interviews are used to ask for data over the past 12 months. The SCF asks for current values of some data (e.g.,current 1983 value of stocks) and for annual figures for other data (e.g., income for calendar year 1982). In 1983, SCF data were collected between February and August; in 1986 telephone interviews were conducted between June and September to collect these data.

The primary unit of collection between the two surveys differs, which may pose a problem in interpreting the data. The CEX is used to collect data from a consumer unit (CU) which is defined as 1) a single person living alone or sharing a household with others but who is financially independent, 2) members of a household related by blood, marriage, adoption, or other legal arrangement, or 3) two or more persons living together who share responsibility for at least two out of three major expenses (ie., food, housing, and other expenses). The unit of collection for the SCF is the primary family in any household; families are broadly defined to include single persons.

The CEX distinguishes between the reference person, the person in whose name the residence is owned or rented, and the respondent (who need not be the reference person). The SCF in 1983 used the concept of head of household, defined to be the male in a married couple family, rather than the reference person. The SCF respondent was identified as the person in the family most financially knowledageable, the person more economically dominant, or the person closest to age 45. These differences between heads and references persons and in the definition of respondents are subtle but may affect the alignment of the data.

Sources of error in both surveys include sampling errors, non-response errors, reporting errors, errors due to differences in interpretation of questions by respondents, and mistakes in recording or coding the data. Because of the different sampling frames, different sampling errors are expected in the two surveys. Nonresponse errors are due to non-interviews, refusal to participate, and incomplete interviews. The response rate for the 1983 CEX was 86%, for 1983 SCF area probability sample the response rate was 71 %, and for the high income 1983 SCF sample it was 95.5%. While these differences are significant, both surveys contain weights which can be used to produce national population totals from the sample data; incorporated in the weights is adjustment for some types of non-response.

Reporting errors are checked by the respective agency staff for internal and external consistency. In some cases, imputation procedures are used to dean the raw data and to estimate missing values. However, values were not imputed in the CEX for missing income information. In contrast, for the 1983 and 1986 SCF data, imputation was used to create values for missing income and other financial data. Both data sets allowed for valid non-responses.

Various weights are available in the two data sets to adjust for non-response and to enable the samples to represent the U.S. population. The SCF included separate weights for inclusion in the "cleaned" data and meshed weights for the combined area probability and high income sample. The last weight listed in the CEX file, FINLWT21, was used for producing population totals.

Both the CEX and the SCF collected significant amounts of demographic data, such as the age, sex, and education of CU/family members, household composition, race and ethnicity, marital status, occupations of CU/family members, labor force participation, region and area type, and hours and weeks worked. We considered the alignment of these variables critical to any matching or merging of the two data sets.

The CEX collected before and after tax income data from CUs. Income measures included earned income, asset income, government income supports (including the cash value of food stamps), retirement income, other income (e.g., alimony, child support, scholarships) and other lump sums. The CEX also collected information for some types of in-kind income (i.e., meals as pay, rent as pay); information on in-kind income was not collected in the SCF. The 1983 SCF collected before tax income from the same general sources as did the CEX However, different sources of income were combined in both surveys so that one-to-one comparability was not possible in some instances. For example, the CEX collected data on workers compensation combined with veterans benefits, and on unemployment. The SCF collected data on unemployment combined with workers compensation; information on veterans benefits was not

total income may be nearly comparable, individual income components are less so. In the 1986 SCF, income sources were not as detailed as in 1983; for example, no questions about asset income were asked.

Information on assets and liabilities was collected using both surveys. Because of the purpose and nature of the survey, the SCF has much more detailed information on assets than does the CEX Again, because of the nature of the phone versus in-person interview, the 1986 SCF has combined some measures which were single variables in 1983 (e.g., money market accounts and certificates of deposit were individual measures in 1983, but combined in 1986; many of the liability measures were also combined in 1986). Asset and credit information was collected only in the fifth and final interview of the CEX

Both data sets are available as public use microdata tapes (PUT). In some cases, however, variables internally available to agencies have been excluded or topcoded in the PUTS. For example, for the CEX the market value of the house and mortgage interest payments are included in the PUT, but the outstanding balance on the mortgage is not. Similarly, many of the credit and liability variables have not been included in the PUT for the CEX. For the 1986 CEX, population size is suppressed for all CUs living in rural areas and for all PSUs in the West; region is coded only for urban CUs. For the SCF, regional/ geographic information is excluded for the 438 highincome observations.

Some of the income and expenditure variables are topcoded on the CEX public use tapes. For example, in 1983 and 1986, detailed income variables are topcoded at $100,000, monthly rent is topcoded at $1,000, and property and medical expenses are topcoded at $200,000. No topcoding was done for the SCF data set.

The SCF data set is organized as a "flat file" with one record per household. The CEX public use tape is organized in four primary files: (1) a family file with one observation per CU, (2) a member file with detailed information on individual CU members, (3) a detailed monthly expenditure file, and (4) a detailed monthly income file. Data in files three and four duplicate some of the information in the family and member files; however, data in the first two files are not necessarily monthly.

After comparing the information in the CEX and SCF Survey Matrix, we decided to use only the 1983 data for the composite-matched file data base. Selecting the 1983 time period meant that we would drop non-MSA rural CUs from the CEX sample. However, the decision seemed justified since the 1983 SCF provides significantly more detailed information than is available in the 1986 survey.

Selection of Variables and Aligned Data File Creation

Two types of variables were selected for building the composite CEX-SCF data base. First, comparable variables, those existing in both surveys, were identified. These were available as matching variables. The second type of variable included those for which we were interested in assigning a value in the data when the variable did not exist, or when we thought values from the other survey could be used to improve the quality of the existing data. Variables of the first type were selected on the basis of how well they matched each other or how easily they could be made to align. In some cases alignment was exact; for example, sex of respondent is unambiguous, as are age and marital status. In all, nearly 40 variables proved to be comparable or could be made to be comparable using member level data in the CEX.

For data matching, it was necessary to reorganize the CEX data as a flat file with one record per consumer unit, creating variables from the four primary files as needed. For example, member level data were used to construct consumer unit level data for the age, education, work status, and occupation of the spouse. Other aligned variables included property values for owned home and the sum of the values of other properties, vehicles, and credit variables (e.g., mortgages for owned home and other properties, current and last year's credit balances owed on store charges, credit ,cards, various financial institution loans, insurance loans, medical bills, and other credit sources). Note that credit information is only available on the internal BLS files; these data are not yet released on PUT.

Variables of the second type were selected based upon their uniqueness to or completeness in the individual surveys. CEX variables first selected for testing included total transportation expenditures and total apparel expenditures. Detailed expenditures for these categories were aggregated for use in the matching process. The SCF asset variables selected were real estate assets, total financial assets, and total liquid assets. FINLWT21 and weight 3016 were selected for use with the CEX and SCF data files, respectively.

Aligned data files were created which cover simi-

lar reference periods. Although the exact time period of the SCF could have been reproduced using the CEX monthly and quarterly data, we decided against this adjustment since, in the future, we plan to match records for additional surveys which have slightly different reference periods than those in the SCF. CEX data were drawn from the interview surveys from the first quarter of 1982 through the second quarter of 1983 (the reference period was from the last quarter in 1981 through the first quarter in 1983). Consumer units which completed four interviews any time during these six quarters and had a fifth interview in either the last quarter of 1982, or in either of the first two quarters of 1983 were included. A sample of 2,450 consumer units, representing a population of 38,520,888 consumer units resulted. Demographic, income, asset, and liability data were taken from the fifth interview. Therefore, income and liability data are for one of three reference periods. These periods are each four quarters in length and start with the last quarter of 1981, the first quarter of 1982, and the second quarter of 1982, respectively. Those consumer units in the second group matched-up exactly with the SCF households in terms of reference period. About one-third of the CEX sample are in this group.

The SCF sample included 3665 cross section and 438 high-income observations, representing a population of 83,917,975 households. The cross section sample alone represented a population of 83,9191,064. The reference period for the SCF income data was the 1982 calendar year. The rest of the data were referenced to February through August of 1983.

Comparative Results from the Aligned CEX and SCF Data Bases

Weighted and unweighted means and distributions are presented for variables common to both the CEX and SCF in this section. This type of analysis is a necessary prerequisite to the actual matching process to insure that the data are likely to be comparable at least across the matching variables. When considering the results, it is a good idea to note the following points. First, in the SCF, all data are imputed in cases of invalid non-response. This is not so in the case of the CEX data analyzed here; for income, assets, and liabilities, missing values (valid or invalid) are treated as zeros. Also, certain of the variables have been created from the CEX data base in order to be more nearly comparable to variables found in the SCF; yet, alignments are still not exact. Also, because the CEX sample used here included only consumer units which participated in four interviews, there may be some biases in the data.

Tables 1 and 2 present both unweighted and weighted percentages of area type by geographic region for the CEX data and the SCF data. It should be noted that the high income sample in the SCF was not included in Table 3 since region and area type data are not provided for this group for reasons of confidentiality. The results indicate that the CEX in 1982 represented large urban areas much more frequently than did the SCF. Nearly 60% of the unweighted CEX sample (nearly 70% of the weighted sample) lived in large urban areas. The SCF had three times more of its

TABLE 1: Percentages of Area Type by Region in the CEX
(Weighted Results in parentheses)
Unweighted sample size=2,450/Weighted sample size=38,520,888

	REGION				
AREA TYPE	Northeast	Midwest	South	West	Total
Large MSAs (over 1.2 million)	15.96 (14.80)	6.41 (20.43)	11.51 (22.57)	13.76 (11.48)	57.63 (69.28)
Other MSAs	6.16 (8.56)	6.78 (6.00)	13.27 (8.83)	6.16 (6.22)	32.37 (29.61)
Not MSA	2.24 (0.07)	3.27 (0.12)	2.82 (0.70)	1.67 (0.22)	10.00 (1.11)
TOTAL	24.37 (23.44)	26.45 (26.55)	27.59 (32.09)	21.59 (17.92)	100.00

TABLE 2: Percentages of Area Type by Region in SCF
(High Income Sample Not Included)
(Weighted results in parentheses)
Unweighted sample size=3,665/Weighted sample size=83,919,054

AREA TYPE	REGION Northeast	North Central	South	West	Total
Large MSAs (over 1 million)	8.73 (10.99)	8.29 (9.897)	6.25 (7.62)	7.72 (10.23)	31.00 (38.73)
Other MSAs	7.45 (5.89)	9.85 (7.68)	13.81 (13.41)	7.04 (7.42)	38.14 (34.40)
Not MSA	3.93 (4.48)	9.58 (7.84)	15.12 (12.48)	2.24 (2.07)	30.86 (26.87)
TOTAL	20.11 (21.36)	27.72 (25.42)	35.17 (33.51)	17.00 (19.71)	100.00

unweighted sample from non-metro areas than did the CEX (the non-MSA rural sample was not included in the CEX sample design in 1982). The results also show that the SCF had a substantially larger part (35%) of its unweighted sample from the southern region of the country than did the CEX (27.5%). Results pertaining to region from the weighted samples are very similar for the two surveys.

Table 3 presents weighted means and standard deviations for continuous demographic variables in both surveys. In the CEX, all persons with two or more years of graduate school were grouped together. For education to be considered a continuous variable, we assumed that individuals in this highest education group all had 17 years of schooling. Therefore, the means on the education variables for the CEX may be slightly biased downwards.

The results presented in Table 3 indicate that the distributions of these demographic variables are similar in the two surveys. We found the weighted means from the two surveys to be statistically significantly different from each other in all cases except for the education of the spouse and family size. The significant difference in the weighted means could be caused as much by the fact that we limited the CEX sample as by any differences between the CEX and SCF.

Table 4 presents weighted frequencies of discrete demographic variables in the CEX and the SCF. One should note here that the variable for the race of the reference person had to be recreated in the CEX to align with categories in the SCF. Hispanics are considered an ethnic group in the CEX, while they are considered a racial group in the SCF. The ethnic origins of consumer unit members are available in the CEX, so that a new race variable was created which would count as Hispanic any reference person who considered him/ herself to be of Hispanic ethnicity. Given that the race variable for the CEX was created from the original race and ethnic background variables, the results for each survey are similar, except for the mix of non-Hispanic blacks and Hispanics. This is most likely due to the way in which the new race variable was created for the CEX data.

The results presented in Table 4 indicate that other weighted percentages are for the most part similar in the two surveys. The SCF seems to have a slightly higher frequency of male headed households than does the CEX; this is likely to be related to the respondent/ reference person definitions cited earlier. The marital status variables also prove to be comparable. The housing tenure frequencies tend to differ. This is most likely due to the fact that the CEX sample used was limited to consumer units which had four interviews. Homeowners would be expected to be more likely to complete four interviews than would renters, and therefore to make up a greater percentage of the population of such

TABLE 3: Weighted Continuous Demographic Variables in the CEX and SCF

	MEAN		STANDARD DEV	
	CEX	SCF	CEX	SCF
Age of Reference Person (Head)*	49.2	46.8	16.9	17.3
Age of Spouse*	45.9	43.8	15.1	15.4
Education of Reference Person (Head)*	12.4	12.2	3.3	3.3
Education of Spouse	12.3	12.2	2.9	2.7
Family Size	2.8	2.7	1.6	1.5
Number of Children*	0.96	0.77	1.2	1.1
Number of Persons Over Age 63*	0.33	0.29	0.63	0.60

* The means were significantly different from each other at the .05 level

TABLE 4: Weighted Percentages for Discrete Demographic Variables in the CEX and SCF

CHARACTERISTICS	CEX	SCF
Race of Reference Persons (Head)		
• White, non-Hispanic	82.3	82.3
• Black, non-Hispanic	10.6	12.9
• Hispanic	5.8	3.7
• Other	1.3	1.1
Sex of Reference Person (Head)		
• Male	68.2	73.7
• Female	31.8	26.3
Marital Status of Reference Person (Head)		
• Married	62.0	60.6
• Separated	2.6	3.9
• Divorced	10.5	11.8
• Widowed	11.8	11.4
• Never Married	13.1	12.4
Housing Tenure		
• Owns	69.3	63.4
• Doesn't own	30.7	36.6

Ongoing and Future Activities

Ongoing and future activities include implementing and evaluating the matching process, making several more matches for the CEX-SCF composite data file, continuing to add additional household surveys to the overall project, and devising statistical procedures which will maximize the structure and composition of the composite records.

The Matching Process

The matching process chosen for our project is a hybrid of linear model-based imputation and matching proce dures, drawing heavily on the work of Rubin (1986, 198 and of little (1982,1986). A linear model has been constructed from each survey to give predicted values for the items unique to that survey (see Table 6 for variables included thus far). Thus $y=xB_1$ is generated using the I units in the CEX, and $z=xB_2$ is generated using the J units in the SCF. Next, y and z are computed for every element in both surveys. The selection of matches z^*_i; for unit i in the CEX will be made by minimizing $(z_{I-zi})^2$ over j=1,...,J, yielding unit j*. Then z^*_i; z^*_I, completing $(y,x)_i$ to give $(y,x,z^*)_{I=}$ $((y,x)_{I,zi^*})$. For unit 1 m the SCF, the analogous procedure will be followed to obtain "completed" unit values of the form $(y^*,x,z)_i=(y^*i(x,z)_i)$, where i* minimizes $(y_i - y_i)^2$

For data sets the size of the CEX and SCF, the matching process cannot be based on predictions for all the data items. For the CEX, a prediction equation for total transportation expenditures and total apparel ex-penditures will be obtained as a regression on the shared demographic variables as shown in Table 6. First estimations are based on the primary variables only, i while subsequent estimations are based on both the primary and secondary variables. For the SCF, a prediction equation is obtained for real estate assets, total financial assets, and total liquid assets, again as a regression on the shared demographic variables. Once the paired unit, f or i* ; are identified, the complete data for that unit are appended to give $(yxz^*)_i$; or $(y^*,xz)_i$ respectively.

For the composite data set, new weights need to be assigned. For the CEX, FINLWT21 is the original weight, w_i; for the SCF, B3016 is the original weight v_i. The algorithms for the construction of the final weights for each survey are not available to compute the hypothetical weight each unit would have been assigned had it been sampled in the alternate survey. Therefore, mated weights are calculated from regression equations including as covariates all the available demographic variables used in the process of arriving at the final weights. For the I units in the CEX, the composite weight is defined by $w_{i=}(w_i^{-1}+v_i^{-1})^{-1}$; and for the J units in the SCF, the composite weight is $w_i=(w_i^{-1}+v_i^{-1})^{-1}$.

As shown by Little(1982,1986) an and Rubin(1986,1987), this procedure has certain optimality properties wherever y I x is independent of z I x. However, this strong assumption of conditional independence cannot be presumed to hold in general The robustness of key analyses to violation of this assumption is being investigated for these two surveys in particular by Sedransk and associates. Results from the matching stage and the evaluation stage will be available in the winter of 1990-91.

Other Surveys

Since the inception of the data comparability project, the participation of representatives from additional surveys and agencies has been anticipated. Several of these representatives have agreed to cooperate with us in the project. Recently, with the assistance of some of these representatives, the Survey Matrix was expanded to include information on SIPP, and the CPS. Future surveys to be included in the data comparability project are the RECS, the AHS, the NFCS, and the NMES.

Implications for Research, Policy, and Education

I One by product of this project deals with federal policies on data collection. If federal surveys could be designed with a set of key questions in common, further development of a merged or matched data file could allow analysis of an expanded set of research questions. A major advantage of the expanded data set available through the merging of two or more federal data sets is the potential improvement in modeling for policy analysis. These data could be used to simulate the effects of potential policies by allowing researchers to incorporate more appropriate variables in their analyses.

For example, the composite CEX-SCF data set could allow researchers to explore the marginal propensity to consume for specific expenditures out of different sources of income. Economists and behavioral decision theorists have hypothesized that "all dollars are created equally," and yet there is anecdotal evidence that people differentiate between different kinds of income. The CEX-SCF composite data should allow the testing of

consumer units than of the total population.

Table 5 compares weighted statistics for selected income, asset, and liability components in the two surveys. Some of the components are difficult to compare because they do not include exactly the same information For instance, the SCF includes certain lump sum payments and capital gains as part of income while the CEX does not. Therefore, we would expect the weighted mean for total income to be higher in the SCF, which it is. Also, the figures for "other income" are not hilly comparable, because the SCF includes some other sources which the CEX does not.

The ranges of the distributions of most of the variables are far greater in the SCF than in the CEX. This is most likely due to the fact that the high income sample of the SCF was included in the sample examined here. Also, with a few exceptions, the SCF data are more heavily skewed than the CEX data. This is also most likely due to the high income sample.

The weighted means on wages and salary are quite close for the two surveys. Results from a t-test indicate that the weighted means on wages and salary are not significantly different from one another (however, note that the skewness coefficient on this variable in the SCF is 7.69; the skewness for a normal distribution is zero). Business and farm income has a different distribution in the SCF than it does in the CEX The two surveys also differ in the distributions of asset and retirement income. The distribution of entitlement income does not differ significantly between the two surveys.

The asset and liability data differ between the two surveys. Among the assets, the weighted means for the value of owned home at least appear to be in the same general range. Other variables have highly skewed distributions in the SCF. For instance, the skewness coefficient on the distribution of revolving charge debt is extremely large at 201.06. Among liabilities, all of the weighted means are significantly different in the two surveys.

TABLE 5: Weighted Statistics for Income and Assets and Liabilities in the CEX and SCF

	Mean		Std. Deviation		Range		Skewness		Kurtosis	
	CEX	SCF	CEX	SCF	CEX	SCF	CEX	SCF	CEX	SCF
Income Components										
Total Income (before taxes)*	22,373.30	26,805.91	19,329.94	43,155.86	401,197	3,420,416	1.66	25.42	12.39	1,291.40
•Wages & Salary	17,303.06	16,660.48	19,024.38	21,127.91	150,000	1,000,000	1.44	7.69	6.24	211.78
•Business & Farm*	733.12	3,488.89	7,293.31	18,512.92	310,000	1,724,199	-1.42	15.54	212.81	519.41
•Asset Income*	1,083.81	2,698.57	3,905.61	20,111.73	101,933	3,531,000	6.13	62.38	63.96	7,008.44
•Retirement Income*	2,594.24	2,320.49	5,211.24	5,231.49	52,406	200,000	2.67	5.69	12.91	90.22
•Entitlement Income	491.45	524.46	1,556.92	1,572.55	20,000	25,000	4.31	4.83	25.16	40.83
•Other*	167.62	504.76	1,393.13	3,558.20	44,299	250,000	21.08	37.31	604.39	2,153.07
Assets										
Value of Own Home*	49,420.71	44,638.45	61,194.32	74,281.92	1,100,000	5,000,000	4.32	10.53	45.57	309.68
Checking Accounts*	893.59	1,423.29	5,376.97	6,126.37	350,000	1,005,200	43.93	50.48	2,687.61	5,899.91
Savings Accounts*	5,760.61	2,303.39	20,799.75	8,352.95	415,000	957,672	9.44	22.69	134.08	1,412.30
U.S. Savings Bonds*	204.43	324.81	2,002.35	3,243.19	74,000	900,000	22.55	146.65	655.27	38,700.61
Liabilities										
Total Consumer Debt*	2,380.93	3,800.59	4,774.46	36,488.20	100,000	7,118,508	5.96	114.92	73.96	18,095.30
•Revolving Charge Debt*	1,054.85	900.71	3,015.58	20,725.99	100,000	6,000,000	13.23	201.06	342.41	50,980.53
•Closed-end Debt*	1,326.07	2,899.88	3,157.14	25,981.67	70,869	5,027,000	5.56	103.73	67.18	16,400.58
Total Real Estate Debt*	10,156.10	14,164.72	24,778.35	57,154.16	462,000	8,485,059	6.62	58.60	92.46	7186.40

* The means were significantly different from each other at the .05 level.

TABLE 6: Variables Selected for Matching

Variable	Consumer Expenditure Survey	Survey of Finances
Dependent	total transportation expeditures total apparel expenditures	real estate assets total financial assets total liquid assets
Independent		
Primary	income quintile area type region age of reference person	
Secondary	education of reference person race of reference person sex of reference person marital status of reference person self-employment status of reference person family size housing tenure number of children number of persons > 64 in unit	

this hypothesis in a way which could not be done with either data set alone. Results from this example study would be particularly interesting for retired households who have a significant amount of unearned income: Do retirees spend pension income differently than Social Security income and differently than earnings in retirement? Results may help policy makers formulate improved pension policies, relating to the portability of pensions from one employer to another or to the security of pension funds. Findings may also point out various incentives for certain types of income among retirees; for example, how would removing the earnings limit from Social Security affect expenditures in general or expenditures for specific items? Policy implications for facilitating employment opportunities among older workers or for Social Security payout formulas could be clarified by this research using these data.

In this example, it would also be possible to study the marginal propensity to consume out of the annualized value of assets: What is the marginal propensity to consume out of interest income versus the annualized value of financial assets? These results would contain implications for educators, via the Cooperative Extension System and other adult education programs, leading to the development, implementation, and targeting of programs to help older persons plan for dissaving in retirement.

The net result of all the policy and educational implications discussed in this particular example would be that households could possibly have a better level of living in retirement if better data were available. If data on access to community services and other community resources mold be added to this data set, we could obtain an even better picture of how households function and how to provide better governmental, community, and educational programs and services to help households maintain or improve their economic wellbeing and security.

Endnotes

1. This paper does not reflect an official position of the Bureau of Labor Statistics, U.S. Department of Labor.

2. Special thanks are extended to the following: Susan Banta, formerly in the Division of Consumer Expenditure Surveys, for computer assistance during the early stages of this project; Wenyu Wang, NSF/ASA Research Associate, for statistical and computer assistance; Kimberly D. Zieschang, Chief of the Division of Price and Index Number Research, for endless discussions concerning the original ideas, visions, and possibilities; Eva Jacobs and Stephanie Shipp for supporting our ideas even when questions loomed on the horizon; and other staff members in the Division of Price and Index Number Research and the Division of Consumer Expenditures Surveys, Bureau of Labor Statistics, U.S. Department of Labor for their support and comments.

References

Achdut, L and Y. Tamir (1985), "Comparative Economic Status of the Retired and Nonretired Elderly," LIS-CEPS Working Paper #5.

Aguilar, R and S. Gustafsson (1987), "The Role of Public Sector Transfers and Income Taxes: An International Comparison," LIS CEPS Working Paper #10.

Alter, Hoist E (1974), "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970, " *Annals of Economic and Social Measurement,* 3(2), 373-397.

Avery, Robert B. and Arthur B. Kennickel (1989), "Measuring Wealth with Survey Data: An Evaluation of the 1983 Survey of Consumer Finances;' *Review of Income and Wealth*, in press.

Budd, Edward C. (1971), "The Creation of a Microdata File for Estimating the Size of the Distribution of Income," *Review of Income* and *Wealth, 317*-333.

Buhmann, B., L. Rainwater, G. Schmaus, and T. Smeeding (1988), "Equivalence Scales, Well-Being, Inequality, and Poverty: Sensitivity Estimates Across Ten Countries Using the Luxembourg Income Study (US) Database," LIS-CEPS Working Paper #17.

Base, Ruben (1988), "Data for Food Demand Estimation: Problems and Sources," University of Wisconsin Agricultural Economics Staff Paper Series, No. 280.

_____(1986), "Is the Structure of the Demand for Food Changing? Implications for Projections," in *Food Demand Analysis: Implications for Future Consumption*, Oral Capps and Benjamin Senauer, eds. Blacksburg, VA: Department of Agricultural Economics.

Buse, Ruben, Thomas Cox, and John Glaze (1986), 'The Nature of Demand for Food, " in Consumer Demand and Welfare: Implications for *Food* and Agricultural *Policy, NCR Research* Publication No. 311, jean Kinsey, ed. St. Paul, MN: University of Minnesota Agricultural Experiment Station (#AD-SB-2718)

Carlson, Steven and Robert Dalrymple (1986), "Food Stamp Participation: A Comparison of SIPP with Administrative Records," SIPP Working Paper Series, No. 8604.

Curtin, Richard, Thomas Juster, and James Morgan (1987), "Survey Estimates of Wealth: An Assessmentof Quality," NBERCoiiEererice on Research in Income and Wealth, Baltimore:, MD.

Greenlees, John S., William S. Reece, and Kimberly D. Zieschang (1982), 'Imputations of Missing Values When the Probability of Response Depends on the Variable Being Imputed," Journal *of* the American Statistical Association, 77(378), 251 261.

Hauser, R and L Fisher (1985), "The Relative Economics Status of One Parent Families in Six Major Countries," LIS-CEPS Working Paper #6.

Kilss, Beth and Fritz Scheuren (1978), "The 1973 CPS-IRS-SSA Enact Match Study," Social Security Bulletin 41, (October), 14-22.

Ku, Leighton and Robert Dalrymple (1985), "Differences Between SIPP and Food and Nutrition Services Program Data on Child Nutrition and WIC Program Participation," S1PP Working Paper Series No. 8707.

Lewin/ICF (1989), "The CPS-SOI-CFX Statistically Matched Data Files: Technical Documentation," U.S. Department of Treasury Technical Report to the Congressional Budget Office.

Little, RJ.A. (1982), "Models for Nonresponse in Sample Surveys," journal *of* the American Statistical Association, 77, 237-250.

_____(1986),"Survey Nonresponse Adjustments,*International Statistical Review.*

Okner, Benjamin A. (1972), "Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File," *Annals of Economic* and Social *Measurement, 1(3*), 325-42.

Paris, *Gerhard (1989*), "Stochasiic Generationof a Synthetic Sample from Marginal Information;" Journal *of Business and Economic Statistics*, forthcoming.

Rosenbaum, Paul R (1989), "Optimal Matching for Observational Studies," *Journal of the American Statistical Association*, 84(404),10241032.

Rubin, Donald B. (1986), "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations," *Journal of Business and Economic Statistics*, 4(1), 87-94.

_____(1987),Multiple Imputation for Nonresponse in Surveys. New York. John Wiley and Sons.

Ruggles, Nancy and Richard Ruggles (1974), "A Strategy for Merging and Matching Microdata Sets," *Annals of Economic and Social Measurement, 3*(2), 353-371.

_____and Edward Wolff (1977), "Merging Microdata: Rationale, Practice and Testing," *Annals of Economic and Social Measurement, 6*(4), 407-28.

Scheuren, Fritz (1985), 'Methodological Issues in Linkage of Multiple Data Bases," Panel on Statistics for an Aging Population, National Academy of Sciences, Committee on National Statistics, 155-178.

_____(1989), "A Comment on The Social Policy Simulation Database and Model: An Example of Survey and Administrative Data Integration," Survey *of Current Business*, (May), 40-41.

Smeeding,Timothy and Lee Rainwater(1988), "Comparative CrossNational Research on Income and Economic Well-Being: The Luxembourg Income Study (LIS)," NEER Conference on Research in Income and Wealth, Washington, D.C., May, 1989.

Wolfson, Michael, Stephen Gribble, Michael Bordt, Brian Murphy, and Geoff Rowe (1989), "The Social Policy Simulation Database and Model: An Example of Survey and Administrative Data Integration," Survey *of Current Business*, (May), 36.40.