# Making Complex Data Available to Users with Simple Tools: The Case of the Survey of Consumer Finances[1]

**Ryan Bledsoe, Federal Reserve Board**
**Ryan Bledsoe, FRB, Mail Stop 153, Washington, DC 20551; ryan.m.bledsoe@frb.gov**

**Key Words: Data access, graphical analysis, software**

This paper discusses the development of new software tools to assist users in working with data from the Survey of Consumer Finances (SCF). Although the SCF has only about 4,400 observations, it has far more variables and the relationships among those variables are often quite complex. Historically, the survey data have been provided to the public in SAS data sets. This mode of storage has been a problem for a number of users who do not have access to SAS or to a means of converting the data sets to another format. In an attempt to satisfy public demand for SCF data, extract files containing the variables of most routine interest were made available in Excel format on the project website. Data in this format also proved difficult for users, largely because of the necessity of using weights in the descriptive analyses that are the principal motivation for this file. Recently developed software written in Visual Basic allows users to create properly weighted tables in the Excel data files using a utility that lists all of the features as simple menus. Two new types of pre-packaged analyses derived using a more complex version of the software are also made available: a set of tables in Excel that replicates and extends official publications with the most recent data and a graphical version of this information in the form of a time series chartbook.

The first section of this paper provides a brief summary of the SCF, covering sample design, data collected, and issues involving nonresponse and variance estimates. The second section reviews the advantages and disadvantages of storing these data files in Excel. The third section discusses the development of three custom functions written in Visual Basic that perform weighted descriptive statistics in Excel. The remaining sections discuss software designed to facilitate descriptive analyses performed by SCF users.

## I. 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 attitudes. Interviews for the most recent survey, the 2001 SCF, were conducted via Computer-Assisted Personal Interviewing (CAPI) by NORC, a research organization at the University of Chicago, between June and December of 2001. The median length interview required approximately 79 minutes, although complicated cases took substantially longer. While most interviews were obtained in-person, about 35 percent were conducted by telephone, generally as an accommodation to respondents' preferences. Data are collected on items that are not always widely distributed in the population (e.g. non-corporate businesses or tax-exempt bonds). In order to provide adequate coverage of such variables and to provide good coverage of broadly distributed characteristics (e.g. home ownership) in the population, the SCF combines two techniques for random sampling. The sample is selected from a dual frame that is composed of a standard, multistage Area-Probability (AP) sample 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 statistical records derived from tax returns. The list sample is stratified on an estimated "wealth index", with higher values selected at a higher sampling rate. These records are made available for this purpose under strict confidentiality rules. The list sample is designed to oversample relatively wealthy families but excludes people mentioned by *Forbes* magazine as the 400 wealthiest in the U.S.

Of the 4,449 completed interviews in the 2001 survey, 2,917 families came from the AP sample and 1,532 came from the list sample. The response rate for the AP sample was about 68 percent. The overall response rate for the list sample was about 30 percent, however the rate was only 10 percent for the part of the list sample containing the very wealthiest families.

Both unit and item nonresponse are important issues for the SCF. Weighting adjustments compensate for nonrespondent households. The adjustments include post-stratification to known, external control totals for age, location, and home ownership. For the list sample, frame data on financial income and the wealth index are also used (see Kennickell and Woodburn [1996]). Multiple imputation is used to deal with item nonresponse data (see Kennickell [1998]); for simplicity, imputations are stored as five replicates ("implicates") of each assigned record.

Both imputation error and sampling error are measurable for the SCF. Estimates of the variance due to imputation are computed using five imputation replicates. Estimates of the variance due to sampling are computed using replication methods where samples are drawn from actual respondent records in such a way that the important dimensions of the original sample design are incorporated. These estimates can then be combined to yield standard errors for analysis (see Kennickell [1999]).

## II. Storing aggregate data in Excel

The software tools discussed in this paper are designed to work with an extract file of SCF data within Microsoft Excel. The public version of the aggregate data set includes 94 summary variables, many at a high level of aggregation over a large number of underlying variables. The internal version of this data set contains 176 summary variables. These variables are generated by the "Bulletin" macro. This macro, which is made available to the public on the project website, is written in SAS code and works with the full SCF data set. These variables are used to calculate most of the descriptive statistics reported in the *Federal Reserve Bulletin* article, "Recent Changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances".

SCF staff and public users' needs were considered before making the decision to store and analyze the extract data sets in Excel. During the 1998 survey cycle, these data sets were stored in SAS and the estimates reported in the *Federal Reserve Bulletin* article "Recent Changes in U.S. Family Finances: Results from the 1998 Survey of Consumer Finances" were calculated with SAS procedures (e.g. Proc Univariate). Two objectives set forth for the 2001 survey cycle were linked. Staff looked to develop better ways of computing and reporting basic data tabulations while making SCF data more accessible to public users. Particularly important needs of staff were a means of tracking changes in basic statistics across survey years and a flexible way of charting these basic time series relationships. Programs developed for internal use could then be provided to public users to simplify their descriptive analyses of SCF data. It was determined that Microsoft Excel was a widely used spreadsheet package that could potentially meet the staff's requirements.

When Excel was chosen, two concerns remained unanswered for internal purposes. First, could Excel perform the analysis required for the *Bulletin* article? If so, could this analysis be performed in a timely manner? The analysis to support the writing of that article required the calculation of approximately 3,000 weighted descriptive statistics, percent change calculations for these estimates across survey years, and the charting of just over 700 time series relationships. In 1998, these descriptive statistics were calculated by SAS procedures, while the percent change calculations and time series charts were sketched out manually by the authors of the *Bulletin* article.

In Excel, percent change calculations are easy to perform, charts can be stored in an individual worksheet or embedded in an existing worksheet, and worksheet functions can be combined to calculate complex statistics. Despite this, Excel fell short of staff's requirements in two important ways. First, there is no simple worksheet function available to calculate weighted medians. To make such a calculation with standard functions would require complex restructuring of the data for each estimate. Second, generating and storing approximately 1,000 weighted means, 1,000 weighted frequencies, and 700 time series charts in Excel using only standard syntax is a practical impossibility. Several steps are required to output a single weighted mean, weighted frequency, or time series chart in Excel. One method for calculating a weighted mean or weighted frequency in Excel combines the worksheet functions SUM and IF in an array formula. For example, a weighted mean of the variable "INCOME" conditioned by the first classification of the age class variable (AGECL) requires the following formula to be typed into a cell:

$$= SUM(IF(AGECL = 1, INCOME * WGT)) /$$
$$SUM(IF(AGECL = 1, WGT))$$

where the names AGECL, INCOME, and WGT represent the group of cells that store the values of the aggregate variables AGECL, INCOME, and WGT, respectively. SUM and IF are worksheet functions. To complete this array formula, ctrl-shift-enter must be keyed, which places brackets around the formula. A more complex formula is required for statistics that condition by a classification variable and only consider families with holdings of the variable of interest. Most estimates in the *Bulletin* article are of this type. An example would be the calculation of the weighted mean of transaction account holdings (LIQ) for families with holdings conditioned by the first education classification (EDCL=1). To compute this estimate, the following formula would have to be typed into a cell and ctrl-shift-enter keyed when the formula was complete as follows:

$$= SUM(IF(EDCL = 1, IF(LIQ <> 0, LIQ * WGT))) /$$
$$SUM(IF(EDCL = 1, IF(LIQ <> 0, WGT)))$$

where the name LIQ and EDCL represent the group of cells that store the values of the aggregate variable LIQ and EDCL, respectively. If the workbook names LIQ, EDCL, and WGT are not set a formula similar to the following would have to be typed:

$$= SUM(IF(AA2:AA22246 = 1, IF(BB2:BB22246 <> 0,$$
$$BB2:BB22246 * BZ2:BZ22246))) /$$
$$SUM(IF(AA2:AA22246 = 1, IF(BB2:BB22246 <> 0,$$
$$BZ2:BZ22246)))$$

where the cells AA2:AA22246, BB2:BB22246, and BZ2:BZ22246 store the values of the aggregate variables EDCL, LIQ, and WGT respectively. Clear as mud, right! And, we have not even started to chart 700 time series relationships with a standardized wizard. Typing in a formula of this length for approximately 2,000 descriptive statistics and maintaining over 700 time series charts with a standardized wizard is clearly not efficient.

Other concerns remained about providing data to public users in Excel format. In particular, it was not clear Excel could perform the descriptive analyses desired by external users in an efficient user-friendly manner. In an attempt to satisfy public demand for SCF data, aggregate data sets were made available in Excel format prior to the 2001 survey cycle. Many users found these data sets difficult to work with largely because of the necessity of using weights in descriptive analyses. As noted earlier, weighted mean and weighted frequency calculations can be computed in Excel by combining worksheet functions in an array formula. However, these concepts are beyond the grasp of many public users of SCF data. Further, it is very difficult to use simple Excel worksheet functions to compute weighted medians, which are often a more robust measure of central tendency than the weighted mean for variables with nonnormal distributions. Since many variables stored in the extract files have nonnormal distributions, weighted median calculations are central to analyses performed on these data. The difficulty of performing descriptive analyses in Excel limited the value of providing SCF data in this format. To increase the value of providing these data sets to public users, it was necessary to design tools to facilitate analyses performed on these data in Excel.

Two steps were necessary before Excel could be used to perform analyses required by SCF data users. First, custom functions would have to be developed to perform weighted median, weighted mean, and weighted frequency calculations. Then Excel's platform would have to be modified to facilitate the use of these functions.

## III. Custom functions

Initially it was uncertain whether Excel could perform the required calculations. However, after a few months of background reading, SCF staff developed functions with Visual Basic code to perform weighted median, weighted mean, and weighted frequency calculations in Excel.

These functions were tested extensively to ensure their accuracy. Estimates computed by these functions were compared to estimates computed by existing SAS procedures. The custom functions performed as expected. When present, negligible differences in estimates could be attributed to differences in rounding techniques. Weighted mean and weighted frequency calculations were also compared to calculations performed by basic array formulas within Excel. These calculations were always identical.

Once developed, it was necessary to modify Excel so that these functions could be called efficiently. For internal purposes, these modifications are performed by procedures called by the "Update" macro. For each survey year, beginning with 1989, a workbook stores the extract data set in a worksheet, 14 *Bulletin* tables are each stored in an individual worksheet, and the Update macro with supporting procedures stored as Visual Basic code. The Update macro calls procedures to calculate approximately 3,000 estimates that are stored in the 14 tables in this workbook. These estimates are maintained either by array formulas or one of the custom functions. A subset of these estimates is included in official publications. These estimates are also provided to the public in two sets of tables that are stored in Excel workbooks. One set of tables uses the most recent internal version of extract data as a source; the other set of tables uses the most recent public version of extract data as a source.

## IV. The Update macro

The "Bulletin" macro, written in SAS code, creates an extract file of data for each survey year, starting with 1989. These data sets are converted to .CSV files with the Proc Export procedure in SAS. Each .CSV file is opened in Excel and saved as a workbook. Once these files have been exported to Excel, the Update macro is run. This macro calls a set of procedures to calculate the estimates stored in the *Bulletin* tables.

The first two procedures called by this macro are "Delete Data" and "Create Data". These procedures refresh the data stored in this workbook preparing the tables in this workbook to be recalculated.

Two techniques are used to calculate the statistics stored in these tables. A small percentage of these estimates are maintained by array formulas. An example is Table 2 of the January 2003 *Bulletin* article where "reasons most important to families for saving" are reported. Estimates maintained by array formulas are recalculated with the updated data when the "Recalculate" procedure is called. This procedure recalculates each formula stored in the active workbook. Other estimates stored in these tables are computed by one of the three functions developed for this project (e.g. weighted median). The "Create Statistics" procedure passes arguments to these functions to calculate the requested statistics. For each table element, the procedure determines the contents to be calculated by referring to the positions in the first row and column corresponding to the element. The first row and column are normally hidden to avoid clutter. The first row element contains the name of the variable of interest (e.g. INCOME). The first column stores information that determines both the calculation requested (e.g. weighted median) and any classification of the result (e.g. a particular age class); formally, this information has three elements: a number indicating the level of the classification variable, followed by the name of the classification variable, followed by a number that indicates the required statistic. This information is passed to the specified custom function. The value of the function is stored in the specific table element.

This workbook allows tables and/or statistics to be added or removed easily. Adding a table takes three steps. First, a worksheet must be added. Then variable names are added to the first row and classification variables are included in the first column with statistic and classification identifiers appended to the appropriate side. Finally, the body of the table (e.g. column and row headers) is included. Once these steps have been taken the Update macro is run to generate the new table. To add statistics to an existing table, a similar set of steps are necessary. The worksheet's first row and column are made visible, variables are added to the top row and/or classification variables with classification and statistic identifiers are added to the first column, and the body of the table (e.g. column and row headers) is included. Once the Update macro runs the new estimates appear.

It is also necessary to identify estimates that are based on ten or fewer actual interviews (50 or fewer implicates). The procedure "Less than 50" determines the number of implicates an estimate is based on. Calculations based on 50 or fewer implicates are highlighted to indicate they may not be robust. These estimates are removed from official publications.

## V. SCF chartbook

The Update macro performs the same set of calculations for each survey year beginning with 1989. This consistency allows staff to consider changes in estimates across survey years. Two of the methods used to look at these changes are simplified by Excel. First, percent changes in estimates across survey years are calculated. Since dollar values are stored in 2001 dollars these calculations represent real changes in estimates over time. Second, time series charts of estimates are plotted across survey years. Weighted median, weighted mean, and weighted frequency estimates calculated by one of the custom functions are included in a line plot for each classification group (including "All families" as its own classification group) that include each of the five survey years. Survey year appears on the X-axis and either dollars or percent appear on the Y-axis. Prior to 2001, staff performed these calculations and sketched out these plots by hand. Using Excel allowed staff to automate these tasks during the 2001 survey cycle. These charts are stored in the SCF chartbook.

The chartbook relies on Visual Basic macros to gather data, name each relationship, and provide functionality to a chart sheet so that these relationships can be viewed. The "Import Data" procedure copies relevant estimates, column headers, and row headers from the workbooks that store the aggregate data sets (Update macro) to a worksheet in the chartbook. The estimates included are those that are computed by one of the three custom functions. This includes estimates stored in tables 1, 3, 5, 6, 8, 9 and 11 of the 2003 *Bulletin* article. Once these data are copied to the chartbook they are sorted so that similar calculations for each survey year appear next to each other. Once sorted, a workbook name is set for each group of cells that form a relationship.

A chart with a drop down box in its upper right hand corner is stored in the chartbook. This drop down box contains the list of relationships stored in this workbook. These relationships are listed in the same order they appear in the *Bulletin* tables. The "Update Chart" procedure is called when a selection is made, which charts the selected relationship. The chartbook is provided to public users in both the interactive Excel form and a PDF format that includes all the individual charts.

## VI. SCF tabling utility

Once these functions were developed and tested, SCF staff looked to package them for public use providing users the ability to perform customized descriptive analyses in Excel with the extract data file. These functions are provided to the public in the SCF tabling utility. This utility was first provided to the public in March of 2003. Three custom menu items allow users to access this form. Menu items are added to the command bar, tools menu, and right click shortcut menu. Clicking one of these menu items calls a procedure that makes the tabling form appear.

This utility has many options. One of three descriptive statistics can be selected: weighted median, weighted mean, or weighted frequency. Users can place as many as three conditions on their estimates. Tables can be written to an existing worksheet or to an entirely new worksheet. The SCF tabling form, which captures the user's choices, is structured with seven frames of options. Once users make their selections from these frames and click OK, procedures calculate the requested statistics and place them in a table in the specified worksheet.

The structure of the table generated by this utility is similar to the hidden structure of the tables maintained by the Update macro. Variables of interest appear in the top row and classification variables and classification identifiers appear in the first column.

The first frame allows users to select where to store the requested table. Users can either elect to store the table in an existing worksheet or in a new worksheet. To output a table to a new worksheet, users click the button next to "New worksheet" and enter a valid Excel worksheet name in a text box. To store a table in an existing worksheet users select the button next to "Existing worksheet". All valid worksheet names (note: tables can not be output to the worksheet that stores the data) are included in a drop down box for users to select from. The table appears in the next available row in the selected worksheet.

The second frame allows users to select one of three statistics: weighted median, weighted mean, or weighted frequency. These are the three weighted statistics for which custom functions were built. The third frame allows users to include "all families" in their calculations or condition their output to include only families with holdings of the variable of interest. Estimates calculated by the weighted median or weighted mean custom functions for the *Bulletin* article only include families with holdings in the calculation, except when INCOME, NETWORTH, or KGTOTAL is the variable of interest. It is assumed that all families have holdings of these variables.

Dollar variables in the public version of the extract data sets are stored in nominal dollars. Thus, all dollar variables in the 1998 extract file are stored in 1998 dollars. The default settings of the tabling utility return dollar estimates in nominal dollars. The fourth frame of the tabling form provides users the option to calculate their tables in real dollars of any one of the five survey years.

Variables of interest are selected from the fifth frame. Users must select at least one variable from the list in this frame, but they may select as many of these variables as they would like to include in their table. All of the aggregate variables stored in the extract are listed in this frame. User-defined variables also appear in this box and may be used in calculations. Selected variables are printed across the top of the table in the order they appear in the list.

The sixth frame allows users to condition their output by classification variables used in the *Bulletin* article or by specific quantiles of any variable stored in the extract data set. There are eight *Bulletin* classification groups to choose from: all families, percentile of income (INCCAT1), age of head (AGECL), education of head (EDCL), race or ethnicity of respondent (RACECL), current work status of head (OCCAT1), housing status (HOUSECL), and percentile of net worth (NWCAT). No limit is set on the number of classification groups that can be chosen. These variables and classifications appear in the first column of the output table. Users may also choose to condition their output by the deciles, quintiles, or quartiles of any of the aggregate variables.

The seventh frame allows users to condition their table on one user-specific condition. This frame includes a list of each of the aggregate variables. Users may select a summary variable to condition their output table by from the list. Once selected, conditions must be set in lower and upper bound text boxes that are provided. The default lower and upper bound conditions are the minimum and maximum value of the selected variable, respectively. When a condition is set in this frame every output statistic in the corresponding table is conditioned by these specifications.

## VII. Summary

In an effort to make SCF data more accessible, an extract file of SCF aggregate data has been provided to public users. When this file was first provided its value was limited due to the difficulty of performing weighted descriptive analyses in Excel. Custom functions that perform weighted median, weighted mean, and weighted frequency calculations were developed by SCF staff. Software tools were created to facilitate public use of these functions within this extract file of data. Software tools were also developed to simplify tracking of time series relationships in the data.

## References

**Aizcorbe, A., A.B. Kennickell, and K.B. Moore** [2003] “Recent Changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances,” *Federal Reserve Bulletin*, v. 89 (January), pp. 1-32

**Bledsoe, R.M.** [2003a] *2001 Survey of Consumer Finances Tabling Utility*. 24 Feb. 2003. Federal Reserve Board of Governors. 1 May 2003 <http://www.federalreserve.gov/pubs/oss/oss2/2001/scftabling2001.zip>

**Bledsoe, R.M.,** [2003b] *Survey of Consumer Finances Chartbook (Interactive).* 24 Feb. 2003. Federal Reserve Board of Governors. 1 May 2003 <http://www.federalreserve.gov/pubs/oss/oss2/2001/scfchartbook.xls>

**Bledsoe, R.M.,** [2003c] *Survey of Consumer Finances Chartbook (Static).* 24 Feb. 2003. Federal Reserve Board of Governors. 1 May 2003 <http://www.federalreserve.gov/pubs/oss/oss2/2001/scfchartbook.pdf>

**Bledsoe, R.M.,** [2003d] *Survey of Consumer Finances Tabling Manual.* 24 Feb. 2003. Federal Reserve Board of Governors. 1 May 2003 <http://www.federalreserve.gov/pubs/oss/oss2/2001/manual.pdf>

**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.

**Kennickell, A. B** [2003] *Survey of Consumer Finances Bulletin Macro.* 24 Feb. 2003. Federal Reserve Board of Governors. 1 May 2003 <http://www.federalreserve.gov/pubs/oss/oss2/2001.bulletin.macro.txt.>

**Kennickell, A.B.** [2002] "Interviewers and Unobserved Data Quality: Evidence from the 2001 Survey of Consumer Finances ," *Proceedings of the Section on Survey Research Methods,* 2002 Annual Meetings of the American Statistical Association, New York, NY (forthcoming).

**Kennickell, A.B.** [1999] "Measuring Data Quality in the 1998 Survey of Consumer Finances ," *Proceedings of the Section on Survey Research Methods,* 1999 Annual Meetings of the American Statistical Association, Baltimore, MD.

**Kennickell, A.B.** [1998] “Multiple Imputation in the

Survey of Consumer Finances," *Proceedings of the Section of Survey Research Methods*, 1998 Annual Meetings of the American Statistical Association, Dallas, TX.

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

**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., M. Starr-McCleur, and B.J. Surrette** [2000] "Recent Changes in Family Finances: Results from the 1998 Survey of Consumer Finances," *Federal Reserve Bulletin*, (January), pp. 1-29.

**Walkenbach, J.** [1999] *Microsoft Excel 2000 Power Programming with VBA*