MULTIPLE IMPUTATION OF THE 1983 AND 1989 WAVES OF THE SCF

Arthur B. Kennickell,
Board of Governors of the Federal Reserve System

Douglas A. McManus
Freddie Mac

October 1994

Presented at the 1994 Annual Meetings of the American Statistical Association, Toronto, Ont.

MULTIPLE IMPUTATION OF THE 1983 AND 1989 WAVES OF THE SCF

Arthur B. Kennickell, Federal Reserve Board, and Douglas A. McManus, Freddie Mac
Arthur B. Kennickell, FRB, Mail Stop 180, Washington, DC 20551, m1abk00@frb.gov

Key Words: Imputation, Panels, SCF

This paper describes the construction of the panel dataset for the 1983-89 waves of the Survey of Consumer Finances (SCF), focussing on multiple imputation of missing data. The existing literature on panel imputation is limited (Camphuis [1993], Little and Su [1989]). In the first section of this paper, we give some background on the design of the 1983-1989 SCF panel. The next section discusses the general sample design issues that lie behind the SCF, and the following section specializes the discussion to the 1983-89 panel. We discuss the construction of the panel dataset and some of the basic issues in data editing. The next section describes our implementation of an application of the FRITZ imputation system, which was originally developed for the 1989 SCF cross-section. Finally, we present some data on the results of the panel imputations.

I. Background on the 1983-89 SCF Panel

In 1983, the first wave of the SCF panel was conducted as a part of a multi-agency effort, led by the Federal Reserve and made possible by the cooperation of Statistics of Income (SOI) at the Internal Revenue Service. Data were collected by the Survey Research Center of the University of Michigan. The survey was designed to gather comprehensive and detailed financial information from a representative sample of U.S. households. The questionnaire was complex and took, on average, about 75 minutes to administer.

The 1983 SCF respondents were reinterviewed in 1986, and again in 1989. The data from the 1983-1986 panel have previously been processed and analyzed (Avery and Kennickell [1991]). However, the 1986 survey is very different from either the 1983 or 1989 waves of the survey. The 1986 survey was much shorter, and in many ways the data quality was inferior to that of the other two surveys. In addition, for most analytic purposes, the major data needs are related to changes between 1983 and 1989. For these reasons, the 1986 data have been used in the work reported here only for bounding imputations and for constructing some summary variables that were asked directly of only some respondents in 1989.

Both the 1983 and 1989 surveys were previously edited and imputed independently using only cross-sectional information.[1] However, this may not be an appropriate treatment if the data are to be used to analyze intertemporal relationships. For example, if we know in one wave of a survey that a household has an income of $1 million, we would need to capture this information in some way in other waves, and this need is independent of the ordering of the reporting of information in time. However, if one must first have "completed" data at each cross-section and panel stage, over time there may be many versions of the "same" data.

II. Sample Design

The sample design for the 1983 survey uses a dual-frame design to address two fundamental problems inherent in measuring wealth. Some components of wealth (for example, holdings of corporate stock) are highly skewed, while others (for example, mortgage debt) are more broadly distributed (Kennickell and Woodburn [1992]). In addition, wealthier households have a higher propensity to refuse participation in surveys (Kennickell and McManus [1993]). If there is no adjustment for this reporting difference, analysis of the survey results will be biased in many cases.

A standard multi-stage area-probability sample with 3665 of the completed cases (a 71 percent response rate) provides good representation of broadly-distributed characteristics. A special list sample designed using a file of individual tax data maintained by SOI (IRS [1990]) improves the precision of estimates of skewed financial variables and enables systematic corrections for unit nonresponse. The list sample was selected in a way that tends to oversample wealthy households. Under an agreement with SOI, each selected list case was mailed a packet containing a letter requesting cooperation with the survey and a postcard to be returned if the person agreed to participate. In 1983 only about 9 percent returned the postcard, but about 95 percent of those who did so were eventually interviewed (438 cases). While the level of nonresponse is high (even by more recent SCF experience), it is important to note that such nonresponse is <u>implicit</u> in most surveys, but usually there is no means of identifying the problem.

III. The Panel Sample

The 1989 wave of the SCF panel is part of a more complicated design. The 1989 survey was an overlapping panel/cross-section based on the

1983 design and on a new cross-section design for 1989 (Heeringa et al. [1993]). From the 1983 sample, 2,845 cases were selected to be interviewed in 1989. Respondents who had not moved since 1983 were treated as both cross-section respondents and panel respondents, and 1983 respondents who had moved were treated as panel respondents only. Where couples in 1983 had divorced or separated, an attempt was made to follow both parts of the original couple. For cost reasons, there was some sub-selection of the 1983 respondents and former partners resulting in a total of 1,479 panel interviews. All list respondents were followed and 361 were interviewed. The area-probability respondents were subsampled by geography. An attempt was made to interview all area respondents who had not moved and 819 of these respondents gave interviews. Of the remaining eligible movers, households with heads aged between 22 and 44 in 1983 were followed with a one-in-four probability, and older respondents were followed with certainty. In all, 299 area mover cases were eventually interviewed. Another 1,664 cases who were from the new cross-section or who lived at 1983 sample addresses have only cross-sectional representation.

The weighted response rates for the area-probability and list panel samples in 1989 was 67 percent and 81 percent respectively. The rate for the area-probability sample seems unusually low for a reinterview. Given the degree of willingness list respondents had to express to be interviewed in 1983, it is not surprising that their response rate is higher. In terms of 1983 characteristics, the area-probability panel respondents tend to be younger than nonrespondents, and to have higher income and wealth. The age difference may partially be explained by the fact that some selected people must have died in the six-year interval, but information was not always available to treat them as ineligible. The income and wealth result probably reflects the fact that wealthier people tend to exhibit more stable residence, and thus are easier to locate. In terms of a few key characteristics, the list sample panel respondents differ only slightly from the entire list sample.

IV. Assembly of the Panel Dataset

Although the 1983 and 1989 SCFs differ somewhat in the set of questions asked, a more serious difference for reimputing the data is the way the data were stored. Unlike the 1989 SCF dataset. which includes "shadow" variables for each survey variable indicating the original status of the variables, the 1983 dataset stores the raw survey data and the edited and imputed data in separate files without an exact variable-to-variable linkage. In processing the 1983 data, adjustments were made to the raw data, largely either to rearrange information in ways that more closely corresponded to the analytic intentions of the questions, or to incorporate information from the questionnaires that was not coded in the raw data. Often it is difficult to determine what was actually imputed in 1983.

Another difficulty in reimputing the 1983 data is the fact that all of the SCF software for systematic editing and imputation was originally developed for the 1989 survey. Imputations in the 1983 SCF were made using an ad hoc regression-based structure that is no longer available, requiring that we build new software for reimputation.

To reduce the reimputation of the 1983 wave to a manageable problem, we reduced the dimension of the 1983 dataset by constructing a set of key summary variables. For example, in the case of checking account balances, the information on individual accounts in 1983 was summarized in one variable. In constructing the working dataset for reimputation, we made an intensive effort to trace the raw data antecedents of the summary variables by comparing values in the final cleaned and imputed dataset with those in the basic raw dataset. Using the information from this search, we created two auxiliary variables. First, for questions involving a dollar amount, a variable was created to contain the reported part of the summary variable to serve as a lower bound in imputation. For example, if a household reported the amounts in only two of three checking accounts, the first shadow variable would contain the sum of the two known balances. However, because of the complex arrangement of the raw data, in a number of cases the values in the edited and imputed dataset could not be associated with a missing or reported value in the raw data. In such cases, we assumed that the value in the imputed dataset was computed or coded from additional information in the questionnaire after the initial coding. The second set of shadow variables summarizes the original "missingness" status of all of the summary variables.

Although our main interest is in the broad outlines of changes between 1983 and 1989, other researchers may need more detailed information about respondents in 1983. In such cases, we recommend that the imputed summary variables be used with other 1983 data to devise satellite imputation programs (or maximum likelihood models) to account for the missing detailed data.

IV. Data Editing

As a result of earlier work, all 1983 interviews and all 1989 cases (panel and cross-section) were already edited and imputed in a cross-sectional sense when the panel processing began. Two principal types of editing problems remain in the panel dataset. First, respondents may classify the same asset, debt, income, or job in different ways in different waves. Unfortunately, there is very little that can be done about this problem except in the case of a narrow range of assets, such as confusion between personal businesses and real estate where one could, in principle, make an educated guess about whether two assets are the same. Second, there may be large swings in the wealth holdings of households based entirely on reporting error, and several cases with particularly large changes in assets were examined for possible errors. However, because the patterns of missing data can be very complex and many changes are possible over a six-year period, it is difficult to perform sophisticated checks. Some questions were asked of respondents about changes in their finances between 1983 and 1989 that might appear to have value for editing, imputation, and analysis. Unfortunately, this information seems to be largely unusable. It appears that many respondents report implausible (or even impossible) changes in assets when asked directly about changes in their finances (Kennickell and Starr-McCluer [1994]).

V. Panel Imputation

Beginning with the 1989 SCF, systematic and reusable software, the FRITZ system, was constructed for cross-section imputation based on multiple imputation and a type of Gibbs sampling (Kennickell [1991]). The procedure is described briefly as follows. The survey data are assumed to have a joint distribution, say $f(x_1,...,x_n)$. Because the form of the distribution is unknown, we take an agnostic approach to modeling the distribution. We would like to express the distribution as an expansion in terms of observable items, including levels, powers, and interaction terms for all variables. However, the number of survey observations is small relative to the desired number of expansion terms. Consequently, restricted forms must be used to stay within the limits of the degrees of freedom, a very important and constraining limitation. Most imputations in the SCF are based on randomized regression-like models that use estimated covariance matrices as sufficient statistics.

In our model, the variables to be imputed are assumed to have a "clique" structure, meaning that variables may depend on a set of variables smaller than the entire range of possible variables and that imputation may take place sequentially (Geman and Geman [1984]). After each imputation is made, the resulting value is taken to be "real" in the succeeding imputations. Each imputation is made multiply, and these imputations are stored in replicates of each case ("implicates"), rather than as multiple outcomes on a single record (Rubin [1987]).

After the entire dataset has been imputed, the resulting "completed" dataset is used to estimate the covariances and other statistics needed for the next iteration of imputations. The main point of the first iteration is to produce reliable starting values, and given the need to inspect the imputations very carefully, only a single imputation is made at this stage. In higher-order iterations in this implementation of the FRITZ model, we make three imputations. In theory, the iteration continues until the process converges. Earlier work suggests that convergence of key statistics occurs very quickly (the 1989 cross-section imputations appear to have converged by the fifth iteration). Iteration is very costly given that one iteration requires about two weeks of computer time and a larger amount of human time to evaluate the output.

It is important to note that the variables that may be missing for a given observation may include some from the list of most powerful likely conditioning variables. In data structures where an ordering may be imposed on the missing data, there are solutions to this problem (Little and Rubin [1987]). However, in the SCF and in many other complex datasets, such ordering is either nonexistent or impractical to achieve. In the SCF it is not far from the truth to assume that every case has a distinct pattern of item nonresponse. To allow for this variety, the FRITZ software accepts the specification of a general list of covariates for the imputation of a given variable, from which it estimates a general moment matrix and subsets the variables for each imputation to include reported or already-imputed values in an individual "regression.".

In principle, imputation in panels is isomorphic to imputation in cross-sections. In panel imputations, some of the x_i in $f(x_1,...,x_n)$, can be taken as variables from additional waves of a survey. However, most theoretical discussions of imputation pay scant attention to the empirical basis of estimation. Generally, it is simply assumed that there is a source of information that is so rich that nothing limits one's ability to estimate, even though missing data problems may be serious. This assumption, reasonable in developing basic theory, is not available to the imputer. The limitations on variables in a panel is more severe than in a cross-section since there are more variables that are potentially informative about missing data. Some variable selection is required. However, with multiple missing data patterns, automatic model selection techniques are not a feasible. However, it is possible to do general investigations using data from prior years to determine which variables are most powerful as was done to a limited degree here (figure 3 and discussion).

Given the complex structure of the 1989 sample, several modeling possibilities are available. One might use only panel cases to model panel behavior, but this approach would discard a great deal of information about the structure of the world in 1989 based on the pure cross-section cases, and equally importantly, it would discard important degrees of freedom.[2] For the imputations reported here, all of the cross-section cases--and all of the multiply-imputed records of those cases--were used for estimation.[3] Given the inclusion of the cross-section cases, we had to decide how to treat the 1983 data that were not collected for these new cases. There are two obvious possibilities: We could treat the 1983 information for these cases as missing data and actually impute it, or we could "dummy out" the 1983 data for the 1989 pure cross-section cases. We make the latter assumption, which in the case of a linear model for variable Y(i,.) for observation j that:

$$Y_{ij}=(panel=1,else=0)*(1+B_i^{83}X_{ij}^{83})+B_i^{89}X_{ij}^{89}+e_{ij},$$

where B(.,i) is a vector of regression betas, X(.,i) is a set of covariates where X(83,i,j) is zero for pure cross-section cases, and e(i,j) is a residual error.

To implement the reimputation of the 1989 and 1983 data, we modified the cross-section software for 1989 to accommodate 1983 data values as conditioning variables, and built new modules for all of the 1983 summary variables.[4] When models became poorly identified, we reduced the maximum number of potential conditioning variables and retained only those variables with a strong effect or prior reasons for the variables to be included. This work, so easily described, accounts for a large part of the processing time.

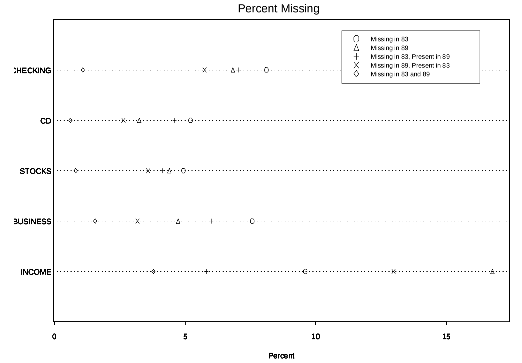
VI. Some Empirical Results

Missing data rates vary widely for variables in 1983 and 1989. Figure 1 shows the unweighted proportion of cases with missing data for total family income and for the components of financial assets. For each variable, the proportion of cases missing the data item in both 1983 is relatively small, a result that should be encouraging if extra-panel data have value in imputation.

Multiple imputation allows us to examine

Figure 1

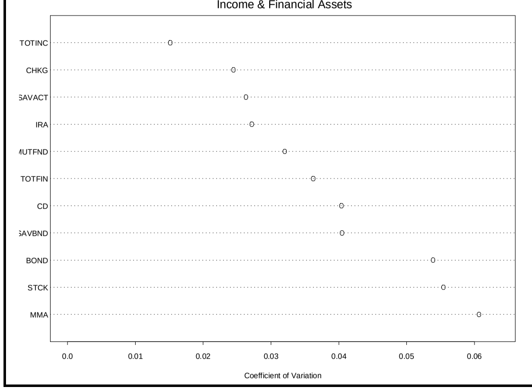
Value of Selected Financial Assets
Percent Missing



how much variability is added to estimates as a result of imputation. Figure 2 shows estimates of the coefficient of variation for the mean of several variables. The means of narrowly-held assets (e.g., bonds) are relatively variable, and those of more aggregated assets (e.g., total financial assets) and more broadly-held assets are less variable.

Figure 2

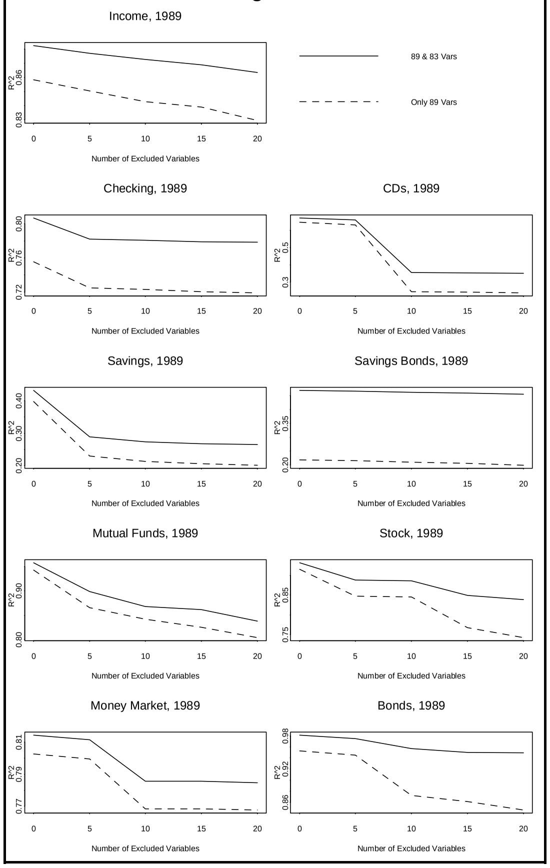
Coefficient of Variation for 1989
Income & Financial Assets



In light of the great variation in the patterns of missing data, an important question is how the performance of the imputation routines degrades as the number of "important" conditioning variables missing for a case increases. A related question is how much effect variables from outside a given wave of the panel have on the imputation of variables within the wave. The results shown in figure 3 provide some information on these questions.

These plots show the decay in R^2 for a series of regressions as variables are dropped. The dependent variables are the logarithms of 1989 total family income and components of financial assets. At the beginning of the series, the explanatory variables include all of the variables that would have been included in the imputation procedure for the variable most similar to these variables. At each

Figure 3



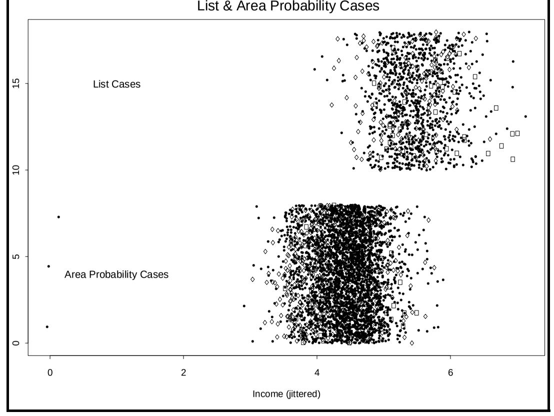
step in the series, we ran a forward-search procedure to identify the five most powerful (in the sense of explaining variance) variables from the maximal set, then we dropped those five variables and re-estimated the model. The charts plot the R^2 of the models against the number of omitted variables. In the upper lines in each plot, the full set of 1983 and 1989 variables was used, and in the lower graph, only 1989 variables were used.[5] Because different cases may have dramatically different amounts and types of missing data, the rate at which the R^2 falls off should indicate how the quality of imputations varies over observations with different amounts of information.

There is considerable variation in the sensitivity of the models to dropping variables. Even after 20 variables are deleted, the R^2 of the regression of total family income falls by only 2 percentage points. In contrast, the R^2 for total savings account balances falls sharply by about 15 percentage points when the first 5 variables are deleted. However, in the range of the usual number of missing data items, the impression is that, the loss of information from dropping variables is small.

As seen by comparing the upper and lower lines in the plots, except for the cases of savings bonds and other types of bonds, the 1983 variables have only a small effect on imputation, probably reflecting the variability of income and the effects of portfolio changes over the six-year period. In the case of savings bonds, the variation is not well-explained in any case. For other bonds, the 1983 variables add about 11 percentage points to the explained variation. Although the data could be taken to suggest that intra-wave information alone is sufficient for imputation, there are three important qualifications. First, in higher-frequency panels, the results of a similar exercise could be very different. Second, some statistical tests turn on such small variations in information that a few percentage points of additional explained variation could reverse the results of a test. Finally, it is also possible that other variables show a higher degree of "persistence." More work is needed here before we can make a clearer judgment.

Ideally, we would like to see the distributions of the imputed data displayed in several dimensions. Although we have made great progress in this regard, we are still largely constrained to look at only bivariate plots. Figure 4 shows the distribution of the reported and imputed data for 1989 total family income for one implicate (•=reported, □=range-card-based imputation, ◊=other imputation). The outliers in the plot derive from values provided by respondents on range cards.

Figure 4
Distribution of 1989 Income
List & Area Probability Cases



A particularly important dimension of variation in the panel is the variation between waves. Figure 5 shows the values for one implicate of total income in 1983 plotted against its value in 1989 (•=reported 83 & 89, ∨=imputed 83 ≺=imputed 89, ◊=imputed 83 & 89). The data cluster about the 45 degree line and the imputations tend to be broadly dispersed over the data cloud, suggesting that the imputations are not inducing large distortions in the longitudinal dimension.

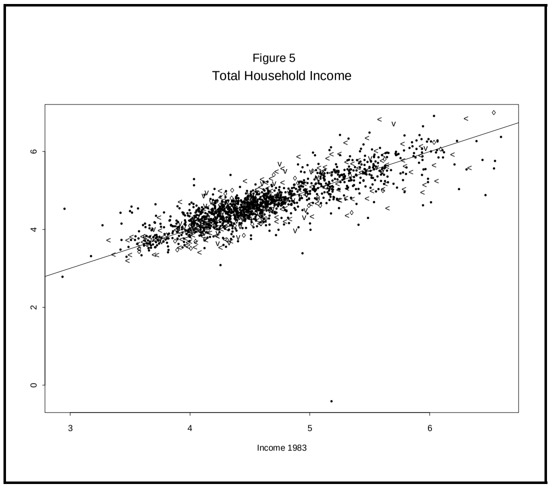


Figure 5
Total Household Income

Bibliography

AVERY, R. B., ELLIEHAUSEN, G. E., and KENNICKELL, A. B. [1988]. "Measuring Wealth with Survey data: An Evaluation of the 1983 Survey of Consumer Finances," Review of Income and Wealth, pp. 339-369.

__________ and KENNICKELL, A. B. [1992]. "Household Saving in the U.S.," *Review of Income and Wealth*, pp. 409-432.

CAMPHUIS, HERMAN, [1993]. "Checking, Editing and Imputation of Wealth Data of the Netherlands Socioeconomic Panel for the Period '87-'89," working paper Tilburg University.

GEMAN, S and GEMAN, D. [1984]. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-6, No. 6 (November), pp. 721-741.

HEERINGA, S., CONNOR, J. and WOODBURN, R. L. [1994]. "The 1989 Survey of Consumer Finances, Sample Design Documentation," working paper, ISR, University of Michigan.

INTERNAL REVENUE SERVICE, [1990]. Individual Income Tax Returns 1987, Department of the Treasury, pp. 13-17.

KENNICKELL, A. B. [1991]. "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," Proceedings of the Section on Survey Research Methods, ASA.

__________ and MCMANUS, D.A. [1993]. "Sampling for Household Financial Characteristics Using Frame Information on Past Income," Proceedings of the Section on Survey Research Methods, ASA.

__________ and SHACK-MARQUEZ, J. [1992]. "Changes in Family Finances from 1983 to 1989: Evidence from the Survey of Consumer Finances," *Federal Reserve Bulletin,* pp. 1-18.

__________ and STARR-MCCLUER, M. [1994]. "Retrospective Reporting of Household Wealth in the 1989 Survey of Consumer Finances," mimeo, Federal Reserve Board.

__________ and WOODBURN, R.L. [1992]. "Estimation of Household Net Worth Using Model-Based and Design-Based Weights: Evidence from the 1989 Survey of Consumer Finances," April 1992, mimeo Federal Reserve Board.

LITTLE, R.J.A. and RUBIN, D. [1987]. *Statistical Analysis with Missing Data*, Wiley, New York.

__________ and SU, H-L. [1989]. "Item Nonresponse in Panel Surveys," in *Panel Surveys*, D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh (eds.), Wiley, New York.

RUBIN, D.B.. [1987]. *Multiple Imputation for Nonresponse in Surveys*, Wiley: New York.

Endnotes

1. See Kennickell and Shack-Marquez [1992] and Avery et al. [1992] for information on the surveys.
2. Among other options, one could include the 1983 cases that were not selected for reinterview, or that were nonrespondents in 1989. Though attractive, this approach is infeasible here because of the editing required to create the summary variables.
3. In calculation of moment matrices, the five imputation replicates of the 1989 pure cross-section cases were down-weighted to account for the multiple inclusion of the same "real" case.
4. Conditioning variables generally include terms to control for the original design, and interviewer observations that we might expect would be correlated with idiosyncratic item nonresponse.
5. We constrained the search procedure to retain the 1983 variables to provide an indication of the largest possible effect of the 1983 variables.

The authors would like to thank Fritz Scheuren for comments and critical support during all phases of the development of the SCF. Thanks also to Herman Camphuis, Steven Heeringa, Dan Skelly, Louise Woodburn, Robert Denk, Gerhard Fries, Janice Shack-Marquez and Martha Starr-McCluer. James Faulkner provided able research assistance for this paper. The opinions presented here are the responsibility of the authors alone and do not necessarily reflect the views of the Board of Governors of the Federal Reserve System or Freddie Mac.