Section VI. Summary of Methodology

A. Specification of FHA Mortgage Termination Models

This Review applies statistical techniques consistent with the literature and applicable to the FHA experience. The purpose of the analysis is to estimate, for FHA loans on the books as of the end of FY 2007, their future probabilities of default and prepayment, so as to compute future outstanding balances, cash flows, and capital ratios. Using loan-level data, ordinary regression analysis breaks down, because the dependent variable indicating default or prepayment is not continuous, but rather is discrete: it is a "1" if either a prepayment or a default occurs in any given quarter or a "0" otherwise (i.e., it is an active loan). One of the problems for ordinary regression analysis in this situation is that the estimated probability of default is not constrained to be between 0 and 1. Binomial logit analysis is one widely used technique to deal with this issue, and has been applied here.

Further complicating the statistical analysis is the fact that mortgage borrowers possess two mutually exclusive options, one to prepay the loan and the other to default on it. From a lender's or insurer's point of view, these are "competing risks" in the sense that they are mutually exclusive, and realization of one risk precludes the other. Prepayment means cessation of cash flow in the form of any mortgage insurance premiums, but thereafter eliminates any chance of default. Conversely, defaulting means default costs are incurred and uncertainty about the possibility and timing of prepayment is elimintated. These competing risks present unique challenges for statistical estimation.

Multinomial logit regression is a general approach to deal with these competing risks, but it is computationally difficult, even for today's high-powered computers. An equivalent technique, based on separate binomial logit estimates for default and prepayment can be used when appropriate adjustments are made for competing risks.   The adjustments for competing risks are of two types: (1) to the data used for estimation, and (2) to the resulting estimated coefficients from the separate binomial logit regressions.

The adjustments needed are to the data used to estimate the equations: for the default equation, eliminating—or "censoring"—the loan's observations in the quarter of the prepayment and subsequently; and for the prepayment equation, eliminating/censoring the observations during the quarter the delinquency starts that leads to a claim and subsequently (during that delinquency/default period, the risk of prepayment becomes zero; if the delinquency is not followed by a claim, the loan remains in the prepayment estimation database).

Once the separate default and prepayment logit equations are estimated, the appropriate multinomial logit probabilities of default and prepayment are computed mathematically from the separate estimates. Another motivation to estimate separate binomial logit models is that FHA

mortgage risk exposure is ultimately determined by the timing and frequency of claim events, rather than the default events that precede insurance claims on the MMI Fund. Thus, the default events (ultimately leading to a claim termination) that result in censoring of prepayment occur with different timing than the claim events. Separating the estimation of logit models of prepayment and claim events simplifies accounting for this timing difference.

Following is an overview of the statistical approach used in this Review. Appendix A provides additional details of these steps, as well as a description of the variables used to "explain" prepayment and claim terminations.

The general approach used in this Review is similar to the multinomial logit models reported by Calhoun and Deng (2002) that were originally developed for application to OFHEO's risk-based capital adequacy test for Fannie Mae and Freddie Mac. The multinomial model recognizes the competing risks nature of prepayment and claim terminations, while the use of quarterly data aligns closely with key economic predictors of mortgage prepayment and claims such as changes in interest rates and housing values.

The multinomial logit models have several benefits over a traditional linear regression. First, it ensures the sum of the three probabilities is equal to 100 percent. This means that at any point in time, a loan can only experience one of the three possible outcomes: prepay, claim, or survive. Second, the possible value of each probability is constrained to be between zero and one under this approach. There is no possibility of estimating a negative probability or a probability exceeding 100 percent. Third, as the probability of one risk increases, the probability of the other risk would automatically be reduced, reflecting the competing-risk nature between prepayment and default. Finally, it allows us to estimate the conditional termination rates using loan-level data. With loan-level observations, the possible outcomes at each point in time are either only 0, the event did not happen, or 1, the event happened. Typical multiple regression models are unsuitable for estimating such discrete dependent variables. Logit regression is specifically designed to handle these types of observations.

Following an approach suggested by Begg and Gray (1984), we estimated separate binomial logit models for prepayment and claim terminations, and then mathematically recombined the parameter estimates to compute the corresponding multinomial logit probabilities. This approach allowed us to account for differences between the timing of FHA claim terminations and the appropriate censoring of potential prepayment outcomes at the onset of default episodes that ultimately lead to claims.

The loan performance analysis was undertaken at the loan level. Through the use of categorical explanatory variables and discrete indexing of mortgage age—in effect classifying loan data into "strata"-- it was possible to achieve considerable efficiency in data storage and estimation. In effect, the data were transformed into synthetic loan pools, but without loss of detail on

individual loan characteristics beyond that implied by the categorization of the explanatory variables. Sampling weights were used to account for differences in the number of identical loans in each loan strata.

Conditional claim and prepayment rates increase quickly during the first two years following mortgage origination before peaking, and then decline slowly over the remaining life of the loan. We applied a series of piece-wise linear spline functions to model the impact of mortgage age on conditional claim and prepayment probabilities. This approach is sufficiently flexible to provide a close fit during the first three years following mortgage origination, including the peak years of claim or prepayment risk, while limiting the number of model parameters that have to be estimated.

B. Differences in the Timing of Borrower Default Episodes and Claim Terminations

For consistency with the available data on loss rates, the incidence and timing of mortgage default-related terminations is defined specifically according to FHA claim incidences. The Begg-Gray method of estimating separate binomial logit models is particularly advantageous in dealing with this requirement. In recognition of the potential censoring of prepayment prior to the actual claim termination date, we used information on the timing of the initiation of default episodes leading to claim terminations to create a prepayment-censoring indicator that was applied when estimating the prepayment-rate model, in effect removing that observation from the prepayment equation database when it was clear from the nature of the delinquency/default/claim path that the probability of prepayment was zero during that time.

Similarly, a separate binomial logit claim-rate model was estimated accounting for censoring of potential claim terminations by observed prepayments, and the two sets of parameter estimates were recombined mathematically according to the above equations to produce the final multinomial model for prepayment and claim probabilities. This approach facilitated unbiased estimation of the prepayment function, which would not be possible in a joint multinomial model of claim and prepayment terminations, since one could not simultaneously censor loans at the onset of default episodes and still retain the observations for estimating subsequent claim termination rates.

To summarize, estimation of the multinomial logit model for prepayment and claim terminations involved the following steps:

- Data on the start of a default episode that ultimately leads to an FHA claim was used to define a default-censoring indicator for prepayment.

- A binomial logit model for conditional prepayment probabilities was estimated using the default-censoring indicator to truncate individual loan event samples at the onset of any default episodes (and all subsequent quarters).

- A binomial logit model for conditional claim probabilities was estimated using observed prepayments to truncate individual loan event samples during the quarter of the prepayment event (and all subsequent quarters).

- The separate sets of binomial logit parameter estimates were recombined mathematically (according to the above equations) to derive the corresponding multinomial logit model for the joint probabilities of prepayment and claim terminations accounting for the competing risks.

C. Loan Event Data

We used loan-level data to reconstruct quarterly loan-event histories by relating mortgage origination information to contemporaneous values of time-dependent factors. In the process of creating quarterly event histories, each loan contributed an additional observed "transition" for every quarter from origination up to and including the period of mortgage termination, or until the last time period of the historical data sample (if the loan remained active). The term "transition" is used here to refer to any period in which a loan remains active or in which claim or prepayment terminations are observed.

The FHA single-family data warehouse records each loan for which insurance was endorsed and includes additional data fields updating the timing of changes in the status of the loan. A dynamic event history sample was constructed from the database of loan originations by creating additional observations for each quarter that the loan was active from the beginning amortization date up to and including the termination date for the loan, or the first quarter of FY 2007 if the loan had not terminated prior to that date.

Additional "future" observations were created for projecting the future performance of loans currently outstanding, and additional future cohorts were created to enable simulation of the performance of future books of business. These aspects of data creation and simulation of future loan performance are discussed in greater detail in Appendix C.

D. Statistical Sampling

The whole population of loan-level data from the FHA single-family data warehouse was extracted for the FY 2007 analysis. This produced a starting sample of approximately 22 million

single-family loans originated between FY 1975 and the first quarter of FY 2007. These data were used to generate loan-level event histories for up to 120 quarters (30 years) of loan life per loan (or until the scheduled age of maturity of the loan).

Estimation and forecasting were undertaken separately for each of the following six FHA mortgage product types: (1) FRM30 – fixed-rate 30-year home purchase mortgages; (2) FRM15 – fixed-rate 15-year home purchase mortgages; (3) ARM – adjustable-rate home purchase mortgages; (4) FRM30_SR – fixed-rate 30-year streamlined refinance mortgages; (5) FRM15_SR – fixed-rate 15-year streamlined refinance mortgages; and (6) ARM_SR – adjustable-rate streamlined refinance mortgages.

We used a 20-percent random sample of FRM30 mortgages and 100-percent samples for all other product types. Loan-level information on borrower FICO scores was introduced for the FY 2007 Review and some historical information on borrower FICO scores was obtained through a choice-based sampling scheme that impacted the sampling of FRM30 loans. Further discussion is provided in the description of loan-level borrower credit score data in Appendix A.

E. Borrower Credit Scores

Borrower credit scores at the loan level are included in the models estimated for the FY 2007 Review. FHA now has relatively complete data on borrower FICO scores for loans originated since FY 2004. In addition, FHA has retroactively obtained borrower credit history information for selected samples of FHA loan applications submitted as far back as FY 1992 and continuing up to FY 2004. These data provide an additional source of loan-level information on borrower FICO scores that were used in estimation. The application of loan-level data on borrower FICO scores is described in greater detail in Appendix A.

F. Cash Flow Model

After the future claim and prepayment rates were projected by the econometric models, the corresponding cash flows were computed. The cash flow computation model includes the calculation of: 1) upfront mortgage insurance premiums, 2) annual mortgage insurance premiums, 3) claim losses, and 4) premium refunds. Two other cash flows were modeled in previous reviews but are not included in our analyses. The administrative expense was discontinued according to Federal credit reform requirements, and distributive shares were suspended in 1990. There is no indication that either of these will be resumed in the foreseeable future. The Federal credit subsidy present value conversion factors published by the Office of Management and Budget are used in discounting these future cash flows to determine their present value as of the end of FY 2007.