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 2005, 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).

Among the problems for ordinary regression analysis in this situation is that the estimated probability of default is not constrained to be between zero and 100 percent. Several techniques are available to deal with this issue, including logit analysis, which is used 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 the lenders' and insurers' point of view, these are "competing risks" in the sense that they are mutually exclusive and one risk, when realized, affects the other. Prepayment means cessation of the mortgage insurance premiums, but zero probability of default thereafter, and defaulting means default costs are incurred but zero probability of prepayment thereafter. These competing risks present a unique challenge 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, binomial logit, when adjustments are made for the competing risks, can be used as the estimation routine separately for each of prepayment and default. 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.

Appendix A provides the detail regarding these steps, as well as a description of the variables used to "explain" default and delinquency. The following is an overview of the the statistical approach used in this Review.

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 starting point for specification of the loan performance models was a multinomial logit model of quarterly conditional probabilities of prepayment and claim terminations. The corresponding mathematical expressions for the conditional probabilities of claims $(p_C(t))$, prepayments $(p_P(t))$, or remaining active $(p_A(t))$ over the time interval from t to $t+1$ are given by:

$$p_C(t) = \frac{e^{a_C + X_C(t) b_C}}{1 + e^{a_C + X_C(t) b_C} + e^{a_P + X_P(t) b_P}} \quad (1)$$

$$p_P(t) = \frac{e^{a_P + X_P(t) b_P}}{1 + e^{a_C + X_C(t) b_C} + e^{a_P + X_P(t) b_P}} \quad (2)$$

$$p_A(t) = \frac{1}{1 + e^{a_C + X_C(t) b_C} + e^{a_P + X_P(t) b_P}} \quad (3)$$

Constant terms a_C and a_P, and coefficient vectors b_C and b_P, are the unknown parameters to be estimated. $X_C(t)$ is the vector of explanatory variables for the conditional probability of a claim termination (versus remaining active), and $X_P(t)$ is the vector of explanatory variables for the conditional probability of prepaying (versus remaining active). Some elements of $X_C(t)$ and $X_P(t)$ are constant over the life of the loan and others are functions of the age of the mortgage.

This specification has 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 remain active. Second, the possible value of each probability is constrained to be between zero and one with 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 risks nature between prepayment and default. Finally, it allows us to estimate the conditional termination rates using

loan-level data. At loan level observations, the possible outcomes at each point in time are only either 0, the event did not happen, or 1, the event happened. Typical multiple regression models are deficient in estimation with such discrete dependent variables. The 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 relatively quickly during the first two years following mortgage origination before peaking and descending more 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 two to three years following mortgage origination, including the peak years of claim or prepayment risk, while limiting the overall number of model parameters that have to be estimated.

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

For the FY 2005 Review, we applied average loss severity rates stratified by mortgage product type. Individual loss severity rates were estimated by using historical average loss severity rates of loans that were claimed during FYs 2000 through 2004 by product type. Differentiation using different LTV ratios was explored but did not show a clear pattern. 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:

1. 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.

2. 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 default episodes (and all subsequent quarters).

3. 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).

4. The separate sets of binomial 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.

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. 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 2005 if the loan has 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. Random Sampling

A full 100-percent sample of loan-level data from the FHA single-family data warehouse was extracted for the FY 2005 analysis. This produced a starting sample of approximately 19 million single-family loans originated between FY 1975 and the second quarter of FY 2005. However, due to data recording delay, the second quarter information of FY 2005 is substantially under-represented. As a result, loans originated during the last quarter of the data extract were only used as reference information and are excluded from the actual analyses. At the estimation stage a 10-percent random sample of loans is used to generate loan-level event histories for up to 120 quarters (30 years) of loan life per loan.

E. Cash Flow Model

After the future claim and prepayment rates are projected by the econometric models, the corresponding cash flows were computed. The cash flow computation model includes the calculation of: 1) upfront mortgage insurance premia, 2) annual mortgage insurance premia, 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 with the FY 2002 Actuarial Review 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. We discount the future cash flows back to the end of FY 2005 by the Federal credit subsidy present value conversion factors to determine the present value of future cash flows.