Annals of Economic and Social Measurement, 1/2, 1972

MICROANALYTIC SIMULATION OF HOUSEHOLD BEHAVIOR*

This paper gives a progress report on the microanalytic model being developed at The Urban Institute for simulating the distributive effects of alternative socioeconomic policies. A sample of individuals and families representing the U.S. population is moved forward in time by a recursive set of functions which predict annual changes in demographic status, earning behavior and wealth accumulation. A macro model of output and price movements provides an environment for the micro model. This paper includes descriptions of the auxiliary macro model, the demographic functions predicting births and marital status changes, and the system of computer programs for developing, implementing and using the substantive models.

1. INTRODUCTION TO THE MICROANALYTIC SIMULATION MODEL

BY HAROLD W. GUTHRIE

Surely the most vexing condition that plagues social science research is the great heterogeneity of the human population. There is great variation between persons with respect to their capabilities for working, their tastes as consumers, and their responses to income changes. Also, a given person, as he proceeds through his life cycle will display many different kinds of behavior over time. Social scientists have faced the tasks of formulating a meaningful description and developing an analytic structure for understanding human behavior by resorting to both extreme abstraction and minute but incomplete details of reality. Given that one function of social science should be to furnish guidance for public policy, the results of social science research have been disappointing. We have offered a mixture of deductive theory too highly simplified to be very relevant to real world issues and inductive empirical findings too remote from a systematic view of a socioeconomic structure.

On its part, government, especially the Federal government, has implemented both macro and micro policy measures on the basis of very scanty information about the expected effects of those policies. For example, we still know very little about the dynamics of an inflationary process and about the dynamics of a deteriorating labor market. With respect to micro policy, Congress has legislated and adminstrators must operate a vast array of special programs aimed at specific subgroups of a heterogeneous population.

One example of a totally chaotic approach to public policy making has been called "American Roulette" and refers to the legislative and administrative processes concerning health care in the United States.[1] Lacking a well-designed plan for provision of health care, a very large number of "health publics," each with its own highly specific self interest, is served by as many as 24 different Federal offices. The resulting combination of a primarily private medical sector and a

* Presented at the NBER Conference on the Role of the Computer in Economic and Social Research in Latin America, Cuernavaca, Mexico, 1971.

[1] Christa Altenstetter, "American Roulette: National Health Policy-Making and Health Programs Implementation," The Urban Institute, Washington, D.C. Working Paper 107-20.

jumble of public programs oriented to special needs produces extreme inequality in the opportunity to live a healthy life. The social and economic consequences of this inequality are issues seldom raised and never systematically analyzed.

What is needed is a dynamic and comprehensive model of our socioeconomic system to guide formulation of public policy. An ideal model would be rich in heterogeneous detail; it would include all of the dominant relationships and variables that describe human behavior; it would allow assessment of a wide variety of social needs; it would be a tool for evaluating the prospective costs and benefits of alternative proposals for changing public policy.

This paper describes a model and a methodology that the authors believe is a first step toward the construction of an ideal model. They are reporting here the results of efforts to build and test a model that encompasses the demography of the U.S. population, as it has changed and as it might be expected to change over time under alternative assumptions about public policy. These efforts are the beginning of a continuing process of expanding the model to be more comprehensive and therefore more useful. The expanded model will include the first stage demographic component reported here, but will also focus, in a second stage, on some important economic issues.

A. Model-building for Economic Policy

Deductive micro economic theory has provided an elegantly logical picture of a heterogeneous world by assuming exact and simple relationships, usually unspecified. This degree of abstraction of the complex real world has led to a model superior to all others yet developed in its comprehensiveness and its ability to reflect an economic system of interdependent activities. Only rarely, however, have data been available to allow specification of the relationships to the degree required for many public policy questions. Even with specification this simplified approach can lead only to single value estimates for large and heterogeneous groups.

The advent of aggregative national accounts data and electronic computers brought a new kind of capability to produce simultaneous equation econometric models. These models focused on fluctuations of the national product in highly specific terms; they have provided useful guidance in the formulation of macroeconomic policy. But the econometric models, even when they are broken down into many sectors and subsectors do not give many insights into the distributional effects of macro policy.

While the importance of macro policy is not to be minimized, the policy needs of our present world are becoming increasingly oriented to micro policy issues. We need to know more than we are now capable of knowing about the distributional dynamics of population change, tax systems, and transfer payment systems. The decade of the 60's brought new awareness of social and economic inequities and proposals of micro policies to reduce the inequities. Given our present methodology in the social sciences, we can only speculate about the long-run effects of these policies. The need, then, is for a method of looking at economic and social processes in a dynamic time dimension but also in as great detail as possible.

As Orcutt has explained elsewhere, changes in state over time for any given variety of elements can easily be conceptualized in terms of a matrix of transitional

probabilities.[2] Given specification of the transitional probabilities, this approach would allow extension of the model through time and would allow microanalysis of distributional effects. The number and the complexity of the matrices required to achieve a comprehensive socioeconomic model not only boggles the mind but would choke even the most advanced electronic computer.

B. Purpose and Design of a Microanalytic Simulation Model

In 1961 Orcutt and his former colleagues at Harvard University, Martin Greenberger, John Korbel, and Alice Rivlin, published the results of their pioneering attempt, started some five years earlier, to build a new kind of model.[3] This model, involving sample representation of decision making units as well as Monte Carlo simulation methods for its solution, laid the foundation for further efforts to provide a basis for policy decisions that would capture the heterogeneity of human behavior.

Three underlying themes form part of the rationale for this kind of simulation model:

(1) The main goal of such efforts is to provide information as input to a decision process leading to the formulation of public policy. While there is methodological ground-breaking to be done, the guiding spirit of the earlier, present, and continuing efforts has as its main focus the fact that wise public policy choices must be made from a wide range of options. The methodology of the model is specifically designed to evaluate alternative social policies by simulation prior to selection of a single policy for implementation.

(2) The output from the model may well have implications for aggregative measures of social well-being, but its principal usefulness lies in its value as a descriptor of the distribution of the population with respect to given characteristics. Thus the model takes as given that the world of social behavior is extremely complex, that the persons and families who make individual decisions vary widely in their states or conditions, and that proper evaluation of social policy must include an examination of its effect on small subgroups of the population. The microanalytic focus of the model attempts to reflect the heterogeneity of the real world.

(3) The model reflects another aspect of reality in that it recognizes that economic behavior and certain elements of social behavior are so closely linked that they should not be separated by traditional disciplinary lines. For example, the welfare position of poor families is determined at least partly by the size of families. Understanding poverty as a social phenomenon therefore requires understanding reproduction as a social process.

The design of the model requires an initial population base at a point in time. The initial population could be arbitrarily designated or it could be a sample representation of the real world. In the demographic model reported here two

[2] Guy H. Orcutt. "Microanalytic Models and Their Solution" in *Mathematical Model Building in Economics and Industry*. London: Charles Griffin & Company Ltd., 1970.

[3] Guy H. Orcutt. Martin Greenberger, John Korbel, and Alice M. Rivlin. *Microanalysis of Socioeconomic Systems: A Simulation Study*, Harper & Row. 1961.

initial populations were used the 1/1,000 sample of the 1960 Census of Population and the 1967 Survey of Economic Opportunity.[4] The initial population is then moved forward through time in annual intervals by imputing changes in the characteristics of the persons or imputing certain events. These imputations are based on relationships that have been discovered through various research efforts in the social sciences. The relationships are referred to as operating characteristics. The imputation proceeds by matching a random number generated within the computer against a calculated probability of occurrence for a given event for a given person: this is the Monte Carlo simulation process. For example, suppose the mortality rate for 85 year old white males is 20 percent. A uniformly distributed random number will be generated for the event, death, for each person of that description in the sample. Any person for whom the random number falls in the bottom 20 percent of the random numbers range will be assumed to have died.

Selection of the operating characteristics is obviously a crucial element in the design of the simulation model. Ideally, the operating characteristics will consist of causal relationships or representations of beliefs and attitudes that motivate human behavior. In the model presented here, the imputation of births best approximates this ideal because the operating characteristics form a sub-model believed to represent the decision processes and other circumstances affecting the occurrence of births. Descriptions of the operating characteristics for births and changes in marital status will be presented here. Operating characteristics for deaths and education have also been developed and they will be described in forthcoming publications. We attempt to validate each operating characteristic as it is developed in order to be assured that we are at least portraying accurately the real world of the past. Currently we are using the 1/1,000 sample of the 1960 Census as an initial population and comparing our annual simulation results with vital statistics for the decade of the 1960's. For the purpose of projecting the future behavior of household, we have to rely on our own best judgment, as well as the insights of other social scientists, about changes in our operating characteristics over time.

Research efforts are under way to expand the microanalytic demographic model to a comprehensive representation of the economic behavior of households. Building upon the demographic base we are well advanced toward specifying operating characteristics concerning labor force participation, weeks worked, annual earnings, receipts of transfer payments, the yield on earning assets held by families, income tax liabilities, disposable income, and saving out of income. We are preparing to be able to project the effects of various manpower programs that are of special interest to the U.S. Office of Economic Opportunity, our current source of funding. We also expect to be able to consider public policy questions relating to transfer payments programs.

Another significant sub-model developed as part of the total simulation modeling effort is an auxiliary model of output and price movements. This model, developed by Orcutt, will complement the micro model by providing an economic

[4] The 1967 Survey of Economic Opportunity, conducted by the U.S. Bureau of the Census for the Office of Economic Opportunity, is a large sample survey which was stratified and weighted to yield a large number of poor persons. For purposes of the simulation model a self-weighting subsample was selected.

environment within which the micro simulation can proceed, and by serving as a calibration device for some of the economic variables in the micro model. The auxiliary model is described in the next section of this paper.

Clearly the development of sub-models and specification of the operating characteristics requires a substantial input of the best products of social science that we can find. But the model would be of little use if we could not implement it with simulation runs on a computer. The design of a simulation system has been a second crucial element in our modeling effort, and George Sadowsky describes the system that he has developed in the concluding section of this paper.

2. THE AUXILIARY MODEL

BY GUY H. ORCUTT

The microanalytic operating characteristics provide the core for our simulation model because the desired output from the model is the assessment of the effects of alternative public policies on the distribution of income and assets. The usefulness of the core model of household behavior can be greatly enhanced, however, by an auxiliary aggregative model that provides closure.

The value of providing closure is two-fold. In the first place the microanalytic models under development need an environment in which to operate. The household sector, after all, does not operate in a vacuum: it is affected by the general condition of the economic system as it is reflected in aggregate unemployment rates, changes in price level, and growth rates. In the second place economists think they know something about the control of some macrovariables such as unemployment rates and changes in price level. It would be useful to trace out the impact of fiscal, monetary, and other policies operated at the macrolevel on the behavior and well being of individuals and families. The macromodel under development represents a first step in providing both an environment for the microanalytic models and a useful link to variables which can be controlled or at least influenced by available monetary and fiscal tools.

The simplest expedient for providing a needed environment for the Urban Institute model of the populations of individuals and families would be to treat unemployment, real GNP, price level changes, and fractions of GNP going to earned income and wealth holders as direct exogenous inputs. The disadvantage of this approach is that no explicit account is taken of the extensive interrelatedness of these variables or of the impact on these variables of what is going on in the microanalytic model. By leaving such variables entirely unconnected the user of the microanalytic model would be given a very unrealistic view of the extent to which outcomes could be independently manipulated by use of policy tools at the macrolevel. The primary objective behind the auxiliary macromodel is to take a useful step towards capturing the close interconnectedness of household inputs from the macrolevel and still leave points at which policy assumptions could be entered either by alteration of target unemployment or by alteration of parameter values. A secondary objective is to provide a macromodel designed to receive

inputs from microanalytic models and so extend the range of application of such models.

In developing an auxiliary macroanalytic model extensive simplification has been achieved by assuming that the federal government can and will cause aggregate demand to vary so as to control approximately the percentage of the labor force which is unemployed. The advantage of this assumption is that if total aggregate demand actually is controlled by the federal government it becomes less critical and possibly unnecessary for present purposes to account for the role of non-household sectors in generating aggregate demand. The behavior of the private sector in this area is simply regarded as being supplemented or offset as necessary to achieve a desired unemployment rate given past price movements. Of course this would not do for a model intended to be useful in guiding short-run stabilization efforts. It is hoped and expected that the model described here will be useful in tracing out the main longer run consequences of monetary and fiscal policy for household behavior and welfare.

In its current stage of development the following relationships form the infra-structure of the macro-model:[5]

1. Target unemployment rates as specified by the user or as a user selected function of change in price level.
2. Unemployment rate as a function of the target rate and a lagged unemployment rate.
3. Labor force exogenously given or as an input from the microanalytic model.
4. Employment as an identity relating to unemployment rate and labor force.
5. Real GNP as a function of lagged GNP, capital, and employment.
6. Real gross private domestic investment as a function of lagged investment, GNP, change in GNP, and change in population given exogenously or as an input from the microanalytic model.
7. Real capital consumption as a function of lagged capital stock.
8. Real capital stock as an identity relating to lagged capital stock, capital consumption, and investment.
9. Implicit price deflator for GNP as a function of lagged price level, unemployment rate and change in unemployment rate.
10. GNP at current prices as an identity relating to price level and real GNP.
11. Capital consumption allowances at current prices as a function of lagged capital consumption allowances and GNP.
12. Net national product as an identity relating to GNP and capital consumption allowances.
13. Indirect business taxes as a function of lagged indirect business taxes and net national product.
14. National income as a function of lagged national income, net national product and indirect business taxes.

[5] A complete description of the initial version of the macromodel which has been computerized is available in Urban Institute Working Paper 504-1, "An Auxiliary Model for Generating Employment, Income, and Price Movements," by Guy H. Orcutt and Sara D. Kelly.

At this stage of its development the auxiliary macromodel should be regarded as a start toward establishing useful links between monetary and fiscal policies and a microanalytic model of the population of individuals and families. It also is of interest in that it provides for and makes important uses of the output of a microanalytic model as input into a macroanalytic model; for example, the labor force and population are estimated from the micro-model.

This model has several deficiencies which hopefully can be reduced with additional effort. Perhaps the most serious of these is that the gap between what policy makers might do at the macrolevel and appropriate alteration of parameter values in this auxiliary model is still uncomfortably large. Also it is unfortunate but true that, while relationships used in this auxiliary model do fit past data very well, important causal relations may not have been successfully captured. In addition, while it may be possible to use fiscal policy and monetary policy to control the level of aggregate demand while also influencing the share of GNP going as a return to wealth holders, this possibility is not explicitly provided for in this auxiliary model as it now stands.

This model has been used to generate a wide variety of outputs including: outputs based on use of observed values as equation inputs; outputs obtained using generated values of endogenous variables as equation inputs; outputs involving a replay of history since 1929 with assumed alterations of policy; conditional predictions for the 1970's assuming alternative unemployment rates and population growth rates; and outputs obtained from sensitivity experiments. Outputs have been generated both with and without suppression of error terms.

The macromodel already developed is considered to be one potential component of a more ambitious and gradually evolving macromodel with several components which interact with each other and with the micromodel. A time series data bank plays a key role in the articulation of macromodel components with each other and with the micromodel. Each model or model component is operated in sequence and when it is being operated can make use of whatever is in the time series data bank as well as add to what is in the bank. Series being generated are distinguished from observed or assumed time series so as to facilitate comparison of generated series with historical observations for purposes of testing, alignment and guidance in seeking improvements. All series in the time series data bank are available for statistical analysis and output is generated in tabular or graphical form.

3. BIRTHS

BY GERALD E. PEABODY

Individual family decisions about the number of children to have are the most important factors affecting fertility in the United States. A couple's desires about the number of children they wish to have is the best predictor of the number of children they will have. Extensive use of contraception has made it possible for a

majority of families to control their fertility to the level they desire. In the Growth of American Families Study, a national survey of fertility conducted in 1955[6] and 1960,[7] it was found in 1960 that 98 percent of the women in the United States intended to use contraception at some point in their life to limit their fertility. Thirty percent of the couples in that study had completely planned their fertility in the sense that all births were planned; each woman conceived only after contraception use was stopped so that she could become pregnant, and she had no unwanted births.

A model that is to explain adequately past fertility or anticipate the probable course of fertility in the future must therefore incorporate considerable detail about family planning. The factors that influence the couple's decision about the number of children they wish to have must also be included along with their attitudes about spacing of children. Their propensity to use contraception when a birth is not desired and the efficacy with which it is used should be analyzed. In addition to these volitional factors, it is necessary to incorporate the physiological capacity of the couple to bear children, and the variation of this capacity over time.

The complete model of the household is being developed as a tool to analyze the potential impact of public policy on the individual and family. Fertility is likely to be affected by changes in attitudes about desired family size, either in response to changes in public policy or other factors affecting attitudes, by changes in the availability of information about existing contraception devices (again, possibly, from new public programs), by changes in the technology of contraception, and so on. By incorporating these features into the fertility model we are well equipped to attempt to anticipate how fertility might respond to such changes effected by public policy or other influences.

A. Theoretical Model

The conventional economic approach to decisions at the family level is to assume a utility function for the family, add a budget and possibly other constraints, and then maximize the utility function subject to these constraints. This approach was initially applied to fertility by Becker[8] some years ago. Becker was led to apply the economic theory of household behavior to fertility since he felt that attempts by psychologists and sociologists had been unsuccessful in accounting for fertility behavior, while economic factors appeared to account for a significant, if small, fraction of fertility behavior. By assuming that children are analogous to consumer durables Becker concluded that couples with high incomes would want larger families than those with low incomes.

While the empirical testing of the economic theory has been scanty, the available evidence indicates that there is very little relation between income and desired number of children. Other economic factors, though, do have a significant

[6] R. Freedman, P. K. Whelpton, and A. A. Campbell, *Family Planning, Sterility and Population Growth*, (New York, 1959).

[7] P. K. Whelpton, A. A. Campbell, and J. E. Patterson, *Fertility and Family Planning in the United States*, (Princeton, 1966).

[8] Gary S. Becker, "An Economic Analysis of Fertility," in *Demographic and Economic Change in Developed Countries*, NBER, (Princeton, 1960).

effect upon fertility; labor force participation of the wife is one such example. However, much of the work on fertility by economists is flawed by the failure to incorporate significant social and physiological factors which are crucial to a complete understanding of fertility.[9] For example, attitudes toward desired fertility are a complex interaction of social, psychological, economic and other causes. Completed fertility is determined not only by these desires, but also by the couple's motivation to meet their desires by controlling excess fertility, and by fecundity factors which are in part outside the control of the couple. Finally, the fertility process occurs over a reasonably long time span during which many of these determining factors may change.

In order to capture this complexity in as much detail as possible, we analyze several components of fertility separately in a recursive framework.[10] No attempt is made to achieve a closed analytic model in which completed fertility is the immediate outcome of the model. Rather, the "solution" of the model is obtained through the cumulative interactions of the equations over the course of the simulation. We first assume that each couple determines the number of children that they wish to have and that they make some decision about the spacing of their children. Not using contraception after marriage or a birth is considered a decision in this framework. It will be further assumed that couples who do not wish to have a child in a given time period will use contraception. Couples will not wish to conceive either because they have already had their desired number of children or because they want to delay the date of their next birth. An effectiveness of contraception use is assigned to those couples who use it. The final factor incorporated into this model is fecundity, the physiological capacity of the couple to conceive and bear a child to full term.

These four aspects of fertility have been incorporated into the following recursive model.

$$N^* = N^*(\text{Dem, SES, Att, N}) \tag{1}$$

$$S^* = S^*(\text{Dem, SES, Att, N}^*, \text{N, S}) \tag{2}$$

$$\text{Eff} = \text{Eff}(\text{Dem, SES, Att, N, N}^*, \text{S, S}^*) \tag{3}$$

$$\text{Fec} = \text{Fec}(\text{Dem, SES, N, N}^*, \text{S}). \tag{4}$$

Here N^* is the number of children the couple wishes to have. It is a function of the couple's demographic attributes, Dem, including their ages, their ages at marriage and their race, and of their socioeconomic status, SES, which includes

[9] See Richard A. Easterlin, "Towards a Socioeconomic Theory of Fertility: Survey of Recent Research of Economic Factors in American Fertility," in S. J. Behrman, L. Corsa, Jr., and R. Freedman (Eds.), *Fertility and Family Planning: A World View*, (Ann Arbor, 1969) for a review of much of the economic literature on fertility (including a survey of the empirical basis of the Becker model) and an examination of some of the shortcomings in the economic theory of fertility as developed at that time. A survey of analytic models constructed by demographers that emphasize physiological factors is contained in M. C. Sheps, J. A. Menken, and A. P. Radick, "Probability Models for Family Building: An Analytic Review," *Demography*, vol. 6 (May, 1969) pp. 161-183.

[10] Further description of this model and the theoretical model underlying it are contained in Gerald E. Peabody, "A Simulation Model of Fertility in the United States," The Urban Institute, Working Paper 709-5, April 30, 1971.

their education, income, labor force status of the wife, and occupation. It may also be a function of the attitudes and values, Att, of the couple and of the number of children, N, the couple now has. Given their desired number of children, we determine the couple's desired minimum interval between children which is a function of the same set of independent variables and in addition may depend upon the desired number of children and some measure, S, of the spacing of their previous children. The effectiveness of contraception use, Eff, is a function of all of these preceding variables. The fecundity of the couple, Fec, is primarily a function of the woman's age, but may also be a function of N and N* (to allow, for example, for the possibility of voluntarily sterilization if $N \geq N^*$) and the other variables indicated.

The number of children that the couple have is then determined by applying this model within the simulation framework. In each year the couple's desired number of children and desired spacing are determined so we know whether or not they desire a birth. Contraception is used or not depending upon whether a birth is not or is desired. The birth probability is a function of the efficacy of contraception, if used, and the couple's fecundity. The Monte Carlo technique is used to determine if the occurrence of a birth is to be imputed. This cycle is repeated for each year of the simulation over the fertile period of the woman, and the total number of children the couple have is determined by the successive outcomes of the Monte Carlo drawings over the total period of the simulation.

By drawing upon a variety of sources it is possible to make reasonable estimates of the equations in the model. Further details of the estimations will be given below, but a brief survey is in order here. The Growth of American Families studies and the National Fertility Surveys[11] provide a twenty year record of families' attitudes toward family planning and their success or lack thereof in carrying out their plans. From these studies, data are available on the desired number of children and the efficacy of contraception use. Additional economic and sociological research and surveys provide other information on the desired number of children. Data on the fecundity of couples, both the distribution among couples of a given age and the decline of fecundity with advancing age, are available from the demographic and biological literature, although it is sometimes sparse. The biggest data gaps are in attitudes about spacing of children. The Princeton Study[12] did ask about spacing desires, but the published results do not link these attitudes to the socioeconomic characteristics of the family.

B. Simulation Model

The term fecundability has been applied by demographers to the monthly probability that a fecund woman will conceive in one month. It is a function of the woman's age and length of the time from her last birth and may be expressed as

$$\text{FEC}(i, t) = \text{F} \cdot f_{it} \cdot \text{F}_1(\text{age}) \cdot \text{F}_2(\text{interval}). \tag{5}$$

[11] Norman B. Ryder and Charles F. Westoff, *Reproduction in the United States: 1965*, (Princeton, 1971).

[12] C. F. Westoff, R. G. Potter, Jr., P. C. Sagi, and E. G. Mishler, *Family Growth in Metropolitan America*, (Princeton, 1961).

Here F is the mean value of fecundability, f_a is a parameter that reflects the distribution of fecundity among women. F_1 gives the age dependence of fecundity, F_2 incorporates the infertile period during a pregnancy and a few months following a birth or miscarriage.

F_1 is zero until a woman reaches puberty, quickly rises to a constant value for the late teens and twenties, and then declines roughly linearly to zero at menopause. F_2 is zero during a pregnancy and for several months after a pregnancy has terminated. For the first two or three months after a birth or miscarriage a woman is sterile, and for several additional months her menstrual cycles are anovulatory. Thus, the period for which a woman is infecund following a pregnancy may be three to five months for a woman who does not nurse and up to a year for one who does.

When a couple is not using contraception, their birth probability, by definition, is their fecundability. Thus for women i in period t

$$\text{PBIRTH}(i, t) = \text{FEC}(i, t). \tag{6}$$

For any couple that does not want to give birth we assume they use contraception. In this model it is assumed that couples for whom $N \geq N^*$ or $S < S^*$ do not want to give birth. For these women the birth probability is given by

$$\text{PBIRTH}(i, t) = \text{FEC}(i, t)[1 - \text{EFF}(i, t)] \tag{7}$$

where EFF is the efficacy of contraception use as given by equation (3).

C. Model Implementation

The first step in the simulation procedure is to assign the variables not originally in the file to women in the initial population. For simulations currently being carried out to align the model, the 1960 Census 1/1000 sample is being used. In this sample no information is available on the desired number of children, desired spacing, or fecundability, so that each of these variables must be imputed. In the initial population, values for these variables are assigned to all married women, and during the course of the simulation they are imputed to women when they marry. Fecundability and effectiveness must be calculated in each year since they depend, respectively, upon the woman's age and whether contraception is used to delay or terminate fertility.

As currently implemented, the desired number of children is expressed as a set of probability functions; i.e., the discrete distribution of N^* is given explicitly. The probability that a couple will desire n children is taken to be given by

$$\begin{aligned} \text{Prob}(N^* = n) &= N_n^*(\text{race, age-at-marriage}) \\ &= a_n + b_n(\text{race}) + c_n(A_m) + d_n(\text{race})(A_m). \end{aligned} \tag{8}$$

Here race and A_m (age-at-marriage) are dummy variables; for example

$$\begin{aligned} \text{race} &= 1, \quad \text{if race} = \text{Negro} \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

A similar relation holds for A_m except that it is a vector of variables corresponding to the different ages-at-marriage that are distinguished. The reasons for this choice of independent variables has been given elsewhere.[13] The Monte Carlo procedure is used to assign a desired number of children to each woman on the basis of this set of probability functions.

One important variable that has been omitted is the labor force status of the woman. Women who want to work and do so have a much lower desired fertility than those who do not work. This variable will be included when the income segments of the model are implemented. Another important variable that has been omitted is the religion of the family. In the U.S. Catholics on average desire almost a full child more than do Protestant families, while Jews desire slightly less children than Protestant families. However, religion is a difficult variable to incorporate into the simulation framework. In the first place it is not available on the census surveys that are being used for the initial populations and so would have to be imputed. More importantly, religion by itself is not the only important factor. The degree of religious committment also has an important influence on fertility, so that we would also have to determine how religious a couple is. The difficulties of making these imputations are so large that this variable is being omitted.

Other variables that are important in determining completed fertility have not been included in equation (8) since they are not very significant for *desired* fertility. Included in this category are education and income. While fertility does vary considerably with education and income, it appears that this variation is a result of the relative effectiveness of contraception practice of different education classes rather than their attitudes about the number of children desired. A final factor that has been omitted for now is the higher fertility that is characteristic of rural farm areas, particularly in the south, relative to urban areas.

The desired minimum interval is also expressed as a series of probability functions:

$$\text{Prob}(S^* = n) = S^*_n(\text{race, education}). \tag{9}$$

These equations have also been estimated in the dummy variable form indicated in equation (8) with two dummy variables for education to distinguish between education of less than high school graduate, high school graduate, and at least some college. Spacing desires are also known to be a function of the desired number of children, but this dependence has not yet been incorporated. We know of no body of survey data available that would enable us to estimate equation (9) directly, so these equations are being estimated by constraining the simulation to replicate the available census data on completed intervals.

Efficacy of contraception use is a function of the variables indicated in equation (10).

$$\text{EFF} = \text{EFF}(\text{race, education, intention}). \tag{10}$$

Intention is the reason for using contraception: either to delay the arrival of the next birth, or to terminate fertility and have no more children. The values of efficacy have been estimated from failure rates for contraception users given in

[13] Gerald E. Peabody, *op. cit.*

the 1965 National Fertility Survey.[14] They range from a low of 0.44 for Negro women who had not completed high school and who wished to delay the next birth to a high of 0.88 for white women with some college education who wished to terminate child bearing. In the absence of empirical data, divorced and widowed women are arbitrarily given efficacy values of 1.0; never-married women are given values of efficacy that will reproduce the rates of births to non-married women.

Fecundability is calculated for each year from equation (5). Since we simulate in intervals of a year rather than a month, the constant F is the probability that a fecund couple will have a live birth in a year rather than the monthly probability of conception. f_{it} is currently solely a sterility index; it is 0 for sterile women and 1 for fecund women. This index is part of each woman's permanent record, although the probability of her being sterile increases with her age. The age dependence in F_1 has been given a simple form that roughly fits the available empirical data. It is taken to be 0 until the age of 17 and then assumes a constant value of 1 until age 28; it then declines linearly to 0 at age 48. F_2 accounts for the fact that a couple that has a birth in one year has a reduced probability of having a birth in the next year due to the reduced exposure time resulting from the infecund period following a birth.

In assigning values to the initial population, more independent variables are required than indicated in equations (8) and (9). Since the married women in the initial population have already had some fertility experience, it is necessary to make the assignment of additional desired number of children conditional upon the current fertility status of the woman. Thus N* is a function of race and the woman's current age, parity and length of marriage. A similar procedure should be followed for the desired interval. However, we are not familiar with any data that would enable us to do so.

Simulations with this model have been conducted with an initial population drawn from the 1960 Census 1/1000 sample consisting of 16,321 individuals who comprise 5,256 households. Simulation runs have been made for a period of ten years, and the resulting simulated birth rates and other fertility measures have been compared with statistics of the actual experience in the 1960s. These initial simulation results are encouraging, and indicate that this model can capture the fertility experience of the decade of the 1960s. Further simulations are currently underway to further improve the model alignment, and detailed simulation results will be made available in forthcoming publications.

4. CHANGES IN MARITAL STATUS

BY STEVEN CALDWELL

Marital status is an important defining characteristic of American adults which assumes an immediate importance for the Urban Institute microanalytic modeling effort because of its close relationship to other personal and family attributes, such as female labor force participation, number and spacing of children, mortality,

[14] N. B. Ryder and C. F. Westoff, *op. cit.*

and family income. To incorporate marital status into the model, transition probabilities must be generated to move persons between the several marital states. One set of generating functions has already been implemented and simulations made; that set will be described below and some simulation results presented. In addition, recent work has improved the conceptualization and estimation of the functions generating these probabilities. Although they have not yet been implemented in an operating model, these newer versions will be described.

A. The Operating Model

Marriage probabilities for never-married persons are generally substantially lower than those for widowed or divorced persons. Thus, for purposes of estimation the population of all single persons is broken into "never-married" and "all others". For the operating model first-marriage probabilities have been made a function of sex, race, and single years of age from 15 through 50. Females have generally higher first marriage probabilities up to age 25 (for whites) and 34 (for non-whites); thereafter males have higher probabilities. In effect, the age profile of transition probabilities both rises and declines more steeply for females. The racial difference in age profiles is also distinct, though perhaps on the decline. Non-white males exhibit a rather flat age profile, which rises fairly rapidly to a peak at a younger age but at a considerably lower level than that of white males and then falls only very slowly for the following 15 years. Roughly the same differences exist between white and non-white females, with the non-white peak being earlier but lower, and the decline considerably slower. Put another way, age seems less useful as a predictor of first marriage decisions for non-whites. Moreover, for both sexes and for almost all ages recent experience reveals non-white nuptiality rates to be substantially lower than white rates.

In the United States, about one-quarter of all marriages in a year are re-marriages. Re-marriage probabilities for those with at least one marriage terminated by death or divorce have been taken from registration data in which the probabilities are tabulated by previous marital status (widowed vs. divorced), age (in three or four broad categories), and sex. Mean values of age-sex specific rates for 1960–1966 were used; no time trend on age-sex-marital status specific parameters was included.

Marriage dissolutions occur in the model through death or divorce. The mortality function, of course, creates widows and widowers. Data on divorce rates reveal that the probability of divorce generally rises to a peak in the third year of marriage and declines thereafter. For the third year and beyond, an exponential function was estimated relating divorce rates to the duration of marriage and used for predicting divorce probabilities for marriages of up to 25 years duration.

Using the above operating characteristics for generating transition probabilities for marriage and divorce, simulation runs have been made on an initial population drawn from the 1960 Census 1/1000 tape. This population was "grown" for 10 annual periods and the observed and expected marriage and divorce rates calculated for various population groups. The results can be seen in Table 1 for marriage and Table 2 for divorce in which the simulated rates are

TABLE 1

HISTORICAL COMPARED TO SIMULATED MARRIAGE RATES IN THE U.S., 1960–69

Year	Total Number of Marriages in 1,000's			Married Rate per 1,000 Unmarried Women over 14			Marriage Rate per 1,000 Unmarried Women, 15–44		
	Historical (H)	Simulated (S)	Percent Deviations[1]	Historical (H)	Simulated (S)	Percent Deviations[1]	Historical (H)	Simulated (S)	Percent Deviations[1]
1960	1523	1490	−2%	73.5	68.9	−6%	148.0	128.3	−13%
1961	1548	1545	0	72.2	71.8	−1	145.4	135.7	−7
1962	1577	1545	−2	71.2	77.8	9	138.4	143.3	4
1963	1654	1755	6	73.4	81.2	11	143.3	148.2	3
1964	1725	1847	7	74.6	81.1	9	146.2	142.1	−3
1965	1800	1831	2	75.0	79.6	6	144.3	139.7	−3
1966	1857	1886	2	75.6	80.0	6	145.1	142.6	−2
1967	1927	2026	5	76.4	82.7	8	145.2	145.6	0
1968	2069	1939	−6	79.1	77.7	−2	147.2	140.5	−5
1969	2145	2038	−5	80.0	82.8	−3	149.1	147.7	−1

[1] $\frac{S - H}{H} \cdot 100$.

TABLE 2

HISTORICAL COMPARED TO SIMULATED DIVORCE RATES IN THE U.S., 1960-69

Year	Total Number of Divorces in 1,000's			Divorce Rate per 1,000 Married Women		
	Historical	Simulated	Percent Deviations	Historical	Simulated	Percent Deviations
1960	393	442	12%	9.2	9.9	8%
1961	414	436	5	9.6	9.8	2
1962	413	435	5	9.4	9.6	2
1963	428	425	−1	9.6	9.3	−3
1964	450	436	−3	10.0	9.2	−8
1965	479	448	−6	10.6	9.4	−11
1966	499	452	−9	10.9	9.3	−15
1967	523	465	−11	11.2	9.4	−16
1968	584	485	−17	12.4	9.6	−23
1969	639	510	−20	13.4	9.8	−27

compared to historical data. Given the small sample size utilized (1,554 persons in the initial sample) it is difficult to draw meaningful conclusions about the discrepancies; however, it does appear likely that the recent rapid increase in divorces beginning in 1967 was not accounted for satisfactorily by the increase in the number of marriages. That is, the rise in the number of low-duration marriages may not be sufficient to account for the increase in divorce; it seems likely that the duration specific rates also changed over time.

B. THE NEWER VERSION

As the model is developed further a richer set of functions to generate probabilities of change in marital status will be incorporated. These newer versions are being developed to help remedy some of the major deficiencies in the operating characteristics described above. These deficiencies are:

(1) lack of time variation in the transition probabilities;

(2) insufficient policy relevance of the independent variables;

(3) important socio-economic differentials omitted.

All three deficiencies are really aspects of the extent to which the existing functions fall short of a true structural model of the marriage and divorce process. With further development and refinement of the operating characteristics for changes in marital status, we hope to capture basic attitudes and behavior that will bring us closer to a structural model and will make it more possible to explore social policy issues concerning marital status.

The distribution of the states of being single (never married), married, separated, divorced and widowed in the national population has long been a matter of public concern, with Americans seeming generally to have considered marriage to be the desirable condition and to have treated other states as, to some degree, "social problems". Thus rising divorce rates have been taken as cause for alarm (although the rising proportion of marriage age persons who are actually

married has received less attention). The value placed on being married is at least partly rooted in evidence showing that, for the U.S., married persons have lower death rates, lower suicide rates, lower usage of facilities for the mentally ill, and probably also lower rates of alcoholism than those widowed, divorced, or never married. It is not clear to what extent, if any, these relationships are causal. But it does seem likely that the possibilities of inter- and intragenerational mobility, both for parents and children, are in general influenced by the distribution and timing of marital transitions. For example, it has been argued that certain of a child's opportunities depend upon whether or not he grows up in a home with both parents.

The new version of the model for first marriage will include a parameter which incorporates education differentials determined from analyses of data in the Survey of Economic Opportunity. Age-sex-race-education specific rates are divided by age-sex-race specific rates, yielding a ratio which is then used to distribute these rates among education classes. This education-specific parameter, however, is not allowed to affect the overall levels of age-race-sex specific rates: a tracking routine scales the education specific ratio up or down to match the predicted age-race-sex rates. Thus, education is, in effect, used as a selective, rather than a causative, factor.

In general, education in moderate amounts seems to increase the marriage rates. Those with a high school education marry at higher rates than the average for nearly all ages. The least attractive partners, (strictly from the point of view of marriage probabilities), especially at middle age or above, seem to be those with the least education followed by those with the most. It has been hypothesized that women "marry up" in socioeconomic status; consequently, high status women and low status men should have the hardest time finding partners and thus have relatively low marriage probabilities. Though such an effect is not easy to locate in available data, college educated women have tended to have lower marriage rates than other women and than college educated males.

Using retrospective data on age at first marriage from the 1960 census, we are now trying to decompose first marriage rates by race and sex into three separate effects:

(1) an age effect—the hypothesis being that there is a constant tendency to marry as a function of age;

(2) a cohort effect —the hypothesis being that all persons born the same year and sharing the same national social atmosphere over their lifetimes have a specific bias about marriage, a greater or lesser enthusiasm for it, which manifests itself in a specific probability at which members of the cohort tend to marry which is more or less than the rate of an average cohort;

(3) a year effect—the hypothesis being that a given year, for economic or other reasons, has a certain depressing or stimulating effect on first marriage rates.

First results from this approach, using parameters estimated from an analysis of variance technique, are promising. The age parameters are all significant and fit a smooth curve with the expected shape. The cohort parameters exhibit an increasing tendency to marry for recent cohorts. And the year parameters seem

to be sensitive to wars and economic trends. We will attempt to use the year effect parameters as a dependent variable with economic outputs from the macro model as independent variables. If successful, this would provide an interesting additional link between the macro and micro models.

The importance of such a decomposition of the independent variables is that it moves us toward a more meaningful policy simulation context by putting more realistic constraints on the ways in which we can vary first marriage probabilities. Furthermore, this does allow time variation in rates, which means we might better capture past experience.

Analyses of data in the Survey of Economic Opportunity allow us to incorporate additional differentials into the re-marriage functions: race, education, and length of time since marriage ended. Again, education is treated as a distributive factor in this function. We also have data for four time periods (1960–1966, 1950–1959, 1940–1949, 1930–1939) so this opens up some possibility of capturing a time trend in these parameters. However, only quite clear trends will be incorporated. The possibility of putting in a time trend simply for simulation purposes is always open.

Given that a set of males and females have "decided to marry," they will be ranked according to their race, education, age and perhaps other variables. The two lists will then be merged to create a marriage in which the partners match as well as possible. Left over males or females will be placed in the lists for the following year. If one sex is consistently in excess, this will in effect replicate the phenomenon of the "marriage squeeze."

Divorce probabilities are increased in sophistication by adding differentials by (1) race, (2) education, and (3) number of children. Number of children could be allowed a causal impact, which might account for some portion of the sharp rise in divorce rates in 1967–1970, since the number of children is inversely related to divorce probability and family size declined over that period.

Finally, we will create a fifth marital status category--separation. We do this for purposes of family income calculation. Separated persons are presently treated as married in our model, but in general we know they must support separate residences, automobiles, etc. Further, the incidence of separation is especially high among non-whites. Thus, to avoid giving an artificially optimistic picture of the non-white population we will create a function which separates marriage partners. Its purpose will simply be to reproduce approximately the incidence of separation in the population.

In the case of divorce or separation, both assets and children must be allocated. We will arbitrarily assign all children to the female. In the case of asset division, we have yet to find any data to guide us.

Some of the most interesting simulation experiments, when the larger model is completed and running, will have to do with the following:

(1) most important, examining the effects of early marriage on fertility, labor force participation, completed education and wealth accumulation;
(2) examining the effects of divorce on subsequent life histories, including the histories of children involved;
(3) and, if it proves fruitful to include in the marriage function year-specific parameters related to the dynamics of the macro economy, examining

the interactions of macroeconomic policy and marriage decisions through such intervening variables as labor force participation and fertility.

5. COMPUTER IMPLEMENTATION

BY GEORGE SADOWSKY

Our microanalytic simulation model of the U.S. household sector is currently being implemented on a PDP-10 computer manufactured by the Digital Equipment Corporation. The PDP-10 computer is a high speed electronic stored program digital computer which operates in a "time sharing" mode. The PDP-10 used for our work is physically located at The Brookings Institution in Washington, D.C., which is approximately one-half mile away from our offices. Its configuration includes 98,304 36-bit words of immediate access memory (of which a maximum of about 70,000 words are available to any one user), 3 demountable disk drives each having a capacity of approximately 5,000,000 words of random access storage, 3 IBM-compatible magnetic tape drives, a punch card reader, a line printer, 5 DECtape drives (a low speed, specialized tape drive), and communications equipment controlling teletypewriter terminals. The computer is accessed by members of the research staffs of several social science research organizations in the Washington area. Access to the computer is obtained through teletypewriters which are connected at the user's end through acoustic couplers to standard voice communication public telephone lines emanating from the computer.

The PDP-10 computer is a "time-sharing" system in the sense that it allows many users concurrent access to its computing resources and that its software is designed primarily for interactive use. Users of such a system typically engage in a "dialogue" with the computer system: a user will type a command to the computer system using a typewriter-like device and will receive a reply typed on the same device indicating the outcome of his request. The computing system may prompt the user to ask for more information during the process of fulfilling the request.

A major difference between this method of using computing machinery, often called interactive or on-line computing, and its historical alternative, batch computing, is that an interactive user can make decisions concerning his research or programming strategy in a sequential manner with assistance from the feedback supplied by the computer, whereas the user of a batch computing system must either prespecify a longer sequence of operations or use more system resources and more of his own time to obtain such flexibility. Interactive computing systems are feasible because of the great disparity between the speeds at which computers and people function and because of the disparate requirements placed upon a computer system by members of its user population. The PDP-10 interactive computing system is analogous to a chess master who can play "simultaneous" games of chess with many human opponents because he can remember

more and think more quickly than his opponents. In the same manner, an interactive computer circulates among its users and allocates its resources to their requirements according to a predetermined system of priorities.

An integrated system of computer programs named MASH (an acronym representing "*M*icro*A*nalytic *S*imulation of *H*ouseholds") is being written for the PDP-10 to help develop, implement and use the family of microanalytic models that will result from the microanalytic research efforts of our staff and of others. The basic unit of simulation within this family of models is the interview unit. As defined by the U.S. Bureau of the Census, it consists of all individuals in a household or other housing unit who are related to each other by blood, marriage or adoption. An interview unit may contain one or more families, and each family may contain one or more persons. Thus, each unit of simulation has a three-level hierarchic structure or tree structure.

Within the MASH system, each interview unit, family and person are assigned names (which are positive integers for programming convenience). For every initial simulation population, the initial set of data describing each interview unit, family, and person is assigned to and is stored in a specific logical address occupied by that interview unit, family, or person. In addition, cross reference information is generated that defines the structure of that initial simulation population. Membership lists generated contain the "names" of all families initially contained within each interview unit and the "names" of all persons initially contained within each family. Address lists generated contain the current logical "address" within computer storage of each interview unit, family, and person in the population. The data for each person include the "name" of the family containing him, and the data for each family include the "name" of the interview unit in which the family is contained. These membership lists, address lists and containment pointers define the structural relationships between entities within the simulation population.

As demographic processes are applied to the simulation population, the initial population structure will change. New names will be assigned to new births, and the data describing the newly born child, including inherited characteristics, will be stored in a new logical person address assigned to the child. Marriages and divorces will generally cause a new family and perhaps also a new interview unit to be created. Deaths will annihilate a person and possibly a family and an interview unit also. For each structural change, the cross reference information is adjusted to reflect the change. In addition, whenever a person changes his family affiliation or creates a new family, the data for the person are moved to a new "address" and the person's address list entry is altered. The person's new family name and a code denoting his reason for leaving his old family are added to the person data at the old address, and the person's old family name and a code denoting his reason for joining his new family are included in the person's data at his new address. After the person has been moved, the data in his old address are preserved indefinitely by the system. Thus, every simulation generates a genealogical record of population structure changes. This information is useful both for programming purposes while building the model and for implementing operating characteristics that require transfers of information among related persons and family units. One use of this genealogical structure is to provide a

mechanism for the inheritance of assets when a family is dissolved due to the deaths of all its members.

An important component of the MASH system is the use of machine readable codebooks for all population data definition and documentation. A MASH codebook is a file of documentation that exists physically as a deck of punch cards or its machine-readable equivalent on magnetic tape or magnetic disk storage. Each sample survey population file that is read by MASH must be defined by such a codebook, and a codebook is automatically generated for every new population file that MASH creates. Each codebook contains: (1) a precise definition of all record types contained in the file; (2) the physical specifications and format layout for each type of record in the file; (3) the unique name, mode, position and label of each attribute (field) in each record; (4) for each attribute, an exhaustive list of values that it can take on and associated labels defining the meaning of each of these values; and (5) sufficient free form text to provide additional file, record and attribute documentation in human readable form. Such a codebook not only provides a unified and complete source of data documentation, but it also allows users of the MASH system to reference any population attributes by specifying the name of the attribute alone.

In addition to providing interactive computing services, the PDP-10 computer system allows its users to maintain on-line random access program and data files of moderately large size. MASH utilizes this feature of the PDP-10 by maintaining its entire current microsimulation population, its address and membership lists, the machine readable codebook describing the population, the user's dictionary of attribute, code and sample definitions, and the time series data bank in on-line random access storage. This form of data organization provides a number of significant advantages for microsimulation modelling. First, the mechanisms for making structural changes within the simulation populations are considerably simpler than they would be within a sequential file processing environment. Second, data browsing functions become quite easy to provide. The MASH user can examine and change any attribute of any entity of the simulation population quickly and at very low cost. Finally, on-line documentation allows the MASH user to refer to attributes solely by name and can provide him with properly labelled output on an interactive basis. The availability of on-line storage devices having substantial capacity is as essential to the viability of the MASH system as it now exists as is the interactive computing environment.

The Fortran IV programming language was chosen as the major implementation language for the MASH system for a number of reasons. Among them were: (1) widespread knowledge and readability of Fortran IV among programmers; (2) efficiency of programming process using a high-level language; (3) ease of interfacing Fortran IV programs with assembly language subroutines; (4) relative ease of exporting and importing programs to and from other computer centers; and (5) existence of an acceptable Fortran IV translator on the PDP-10. Some PDP-10 assembly language subprograms have been added for reasons of efficiency.

The MASH system is designed to be used at different levels for different purposes. For the research user there exists a free form, high level, interactive control language that allows him to create initial simulation populations, control a simulation process based on any one model, examine his population data in

any sequence of his choosing, take censuses of his population and perform a variety of statistical analyses. The existence of this language reflects a belief that productive research is encouraged in an environment that provides a researcher with computing tools sufficiently powerful for him to be in direct contact with and control over his computing activities.

Unfortunately it would be prohibitively expensive to expand such a control language to contain the complexity required for a general microsimulation language, and primarily for this reason the programming implementation of the model's operating characteristics is done in Fortran IV, with some assistance inherent in the MASH system structure. Once the MASH system is complete, we expect that a custodial programmer will be associated with the system to perform the programming required by the inevitable extension, alteration, and maintenance of the model and the education functions associated with an ongoing computer-intensive research project.

The present repertoire of MASH commands may be categorized in the following functional areas: (1) entering and modifying definitions, displaying user defined entities, and other "housekeeping" chores; (2) creating initial simulation populations; (3) data browsing; (4) microsimulation control and execution; (5) taking censuses and obtaining statistical outputs; and (6) adding to, modifying and displaying sections of time series data bank and performing aggregate statistical analysis. The MASH system is organized internally as a modular interpreter, and this form of organization allows us to add to the command repertoire as our experience with the system grows and as we evolve new computing tasks for it. We also intend to modify the syntax of existing commands so that it parallels as closely as possible the language and concepts of social science research.

The scope and character of this high level, interactive command language and some functions are best displayed for the purposes of this paper by a hypothetical computer run using MASH. The example below is indicative of the command language, but does not encompass the entire set of commands.

Let us suppose that a researcher named Gomez wishes to perform a microsimulation of the household sector using a model which contains a known set of operating characteristics. In general, most of these operating characteristics will be embedded within the MASH program, although for the purposes of policy experimentation one or two new ones may have been specified by the researcher and added to the system by the custodial programmer. Further suppose that his population data source is a sample survey file named SEO67 for which there is a machine readable codebook named COD67, and that the attributes in the codebook include those in the following table as well as others:

Level	*Attribute Name*	*Description*
Interview unit	REGION	Region of residence
	NUMFAM	Number of families in interview unit
	URBAN	An urban/rural code
Family	TOTINCOME	Total family income
	NUMPERSONS	Number of persons in family
	ASSETS	Amount of family assets

	DEBTS	Amount of family debt
	FARMVALUE	Value of farm, if any exists
	TAXESPAID	Federal taxes paid by the family
	PENSIONS	Value of private pensions received by family
	SOCSEC	Value of public pensions received by family
	NETWORTH	Net worth of family
Person	AGE	Age of person
	RACE	Code for race of person
	SEX	Code for sex of person
	WAGES	Yearly wages received by person
	WEEKS	No. of weeks person worked
	MARRY	Code for marital status of person
	JOB	Occupation code for person
	HIGHGRADE	Highest grade of school completed

To initiate his simulation activity, the researcher sits at his computer console, dials the computer and makes the connection between the two. He obtains access to the system by entering his account number and his confidential "password." He then enters MASH by typing:

RUN MASH

Each user of the MASH system has his own dictionary which may be used to store variable definitions, recodes, commonly used commands and other system entities. A user would often initially instruct the system to use his dictionary in the event he might want to retrieve from or store into it. To do this, he types:

USE DICTIONARY BELONGING TO GOMEZ:

In order to perform a microsimulation, an initial population must be constructed. The user may include in this population only those attributes and those simulation units that he specifies. Simulation populations are identified by number, and the user declares his intention to describe one by typing:

DESCRIBE POPULATION NUMBER 71:

Information about where to obtain the data for this initial population and how to interpret it are transmitted to MASH in the statement:

EXTRACT FROM SURVEY FILE SEO67 ON UNIT 20
DESCRIBED BY CODEBOOK COD67 ON UNIT 21:

The attributes to be included in the initial population are then specified in one or more statements of the following type:

INCLUDE SURVEY ATTRIBUTES REGION, NUMFAM:

INCLUDE SURVEY ATTRIBUTES TOTINCOME, ASSETS,
DEBTS, NETWORTH, SOCSEC, PENSIONS,
NUMPERSONS, FARMVALUE, TAXESPAID;

Since typing lists of names repeatedly is time consuming, the user is given the option of defining a list of names and then referencing the attributes indirectly

through the name of the list. The list will be stored in the user's dictionary as he has defined it, and can be referenced by name by him during subsequent runs. For example the following statements include all the person attributes listed above in the initial simulation population.

```
DEFINE LIST LABORDATA AS WEEKS, JOB, WAGES;
DEFINE LIST DEMOGRAPHY AS MARRY, AGE, RACE, SEX;
INCLUDE SURVEY ATTRIBUTE LIST DEMOGRAPHY;
INCLUDE SURVEY ATTRIBUTE LIST LABORDATA;
INCLUDE SURVEY ATTRIBUTE HIGHGRADE;
```

Some attributes are not defined within the original sample survey data and must be imputed to simulation units as the initial simulation population is created. The computer instructions to perform the imputation are similar to those that define an operating characteristic, and they have already been added to the system by the custodial programmer. The documentation that describes this attribute has previously been entered in another machine readable file, the attribute library. This library contains attribute definitions for newly created attributes in much the same way that the codebook describing a file contains attribute definitions for attributes whose values are recorded within that file. An example of such an attribute is the number of children desired by a family; it is an important variable for determining the probability of occurrence of a birth. Such attributes are included in the collection of attributes for the initial population by executing a statement of the form:

```
INCLUDE LIBRARY ATTRIBUTE KIDSWANTED;
```

MASH includes a facility for generating a time series of values for an attribute at an individual unit level. For example, suppose it is desired to observe peak net worth achieved and taxes paid by families during the course of the simulation, and in addition the most recent values of public and private pensions received for the last five years. The following commands achieve this:

```
GENERATE HIGHEST 3 YEAR SERIES FOR
NETWORTH, TAXESPAID;

GENERATE LAST 5 YEAR SERIES FOR PENSIONS, SOCSEC;
```

For each series to be generated, a sequence of new attributes is generated, e.g. NETWORTH01, NETWORTH02, NETWORTH03, for each entity in the population having NETWORTH as an attribute, i.e. all families. These new attributes are initially undefined. As the simulation progresses, sequential values of the attribute in time are considered for retention as values of the generated attributes according to the criterion specified in the GENERATE command. Thus, for example, if this population were used as the basis for a simulation of three years or more, then after the simulation was finished *each* family would have values of the attributes NETWORTH01, NETWORTH02, and NETWORTH03 corresponding to the highest values of the attribute NETWORTH generated by

the simulation for that family. Using this command, time series may be generated on an individual basis for any set of attributes at all levels of the hierarchical simulation unit.

After all desired attributes have been specified for inclusion in the population and all time series have been described, the initial population can be drawn. The command to initiate this action is:

```
CREATE POPULATION;
```

Each interview unit in the self-weighting sample survey file SEO67 will appear in the initial simulation population identified by the number 71; i.e., no subselection of interview units will have taken place. If subselection were desired, it could easily be accomplished by defining a *sample* of interview units and extracting only those interview units that satisfied the sample definition. For example, suppose that only single family interview units living in urban areas (URBAN code of 2) were desired. The following command defines a sample of interview units named URBAN1 and stores the definition in the user's dictionary:

```
DEFINE SAMPLE URBAN1 OF INTUNIT AS NUMFAM = 1
.AND. URBAN = 2;
```

Then the command:

```
CREATE POPULATION USING SAMPLE URBAN1;
```

will restrict the initial simulation population to only those interview units that satisfy the definition of URBAN1.

Data browsing commands allow a MASH user to examine the contents of a newly created population or the results of a partial or total simulation. For example, suppose it is desired to search for an old person and examine his pension status and family structure. The following dialogue might take place (MASH responses are underlined):

```
FIND FIRST PERSON WITH AGE >= 65;
PERSON 23 FOUND
EXHIBIT STRUCTURE OF PERSON 23;
INTUNIT 6 CONTAINS 1 FAMILY
    FAMILY 7 LIVES AT ADDRESS 8 AND CONTAINS 3 PERSONS
        PERSON 23 LIVES AT ADDRESS 23
        PERSON 24 LIVES AT ADDRESS 24
        PERSON 25 LIVES AT ADDRESS 25
WHAT IS AGE OF PERSON 23?:
74 . . . AGE IN YEARS
WHAT IS AGE OF PERSON 24?:
68 . . . AGE IN YEARS
WHAT IS AGE OF PERSON 25?:
44 . . . AGE IN YEARS
```

LOOK AT FAMILY 7:

WHAT ARE PENSIONS?:
1760 . . . VALUE IN DOLLARS

WHAT IS SOCSEC?:
825 . . . VALUE IN DOLLARS

WHAT ARE WAGES OF PERSON 23?:
0 . . . VALUE IN DOLLARS

WHAT ARE WAGES OF PERSON 24?:
520 . . . VALUE IN DOLLARS

WHAT IS MARRY?:
2 . . . MARRIED, SPOUSE PRESENT

WHAT ARE WAGES OF PERSON 25?:
2350 . . . VALUE IN DOLLARS

FIND NEXT PERSON AFTER PERSON 24 WITH AGE > = 65:
PERSON 48 FOUND

. . . .

Simulation control commands allow the MASH user to proceed with his simulation run in an incremental manner. For example. to advance the simulation population forward one year in time. executing the following command will suffice:

SIMULATE FOR 1 YEAR USING POPULATION 71:

Suppose that it is desired to observe the effect of the simulation on people who are at least 65 years old. The results of the previous browsing commands can be combined with the incremental simulation commands to halt the simulation for further browsing. For example:

PAUSE AT INTUNIT 6:
SIMULATE FOR 1 YEAR USING POPULATION 71:
AT INTUNIT 6

WHAT ARE PENSIONS OF FAMILY 7?:
1760 . . . VALUE IN DOLLARS

WHAT IS NETWORTH?:
11750 . . . VALUE IN DOLLARS

PAUSE AFTER INTUNIT 6:
CONTINUE:
AFTER INTUNIT 6

WHAT ARE PENSIONS?:
1842 . . . VALUE IN DOLLARS

```
WHAT IS NETWORTH?;
10900 . . . VALUE IN DOLLARS

PAUSE AT INTUNIT 11;
CONTINUE;

. . . .
```

The browsing and simulation control commands may be interspersed to provide on-line control of any simulation. If it is desired, original or calculated values of any attribute may be changed by using the CHANGE command:

```
CHANGE PENSIONS OF FAMILY 7 TO 1800;
. . . VALUE IN DOLLARS

CHANGE NETWORTH TO 11708;
. . . VALUE IN DOLLARS

CONTINUE;

. . . .
```

Output is obtained by the user in two forms: (1) sample surveys of the simulated population; and (2) tabular and statistical outputs. Sample surveys allow the MASH user to extract from a simulation population a new data file, defined by an accompanying system generated codebook, containing only those attributes and those observations the user wants. For example, suppose it is desired to obtain from a simulated population asset, income, and tax data for all families that paid some Federal tax. The following MASH commands:

```
DEFINE SAMPLE TAXPAYERS OF FAMILY AS TAXESPAID > 0;

CONDUCT SURVEY TAXED OF FAMILIES OBTAINING
NUMPERSONS, ASSETS, TOTINCOME, NETWORTH, TAXESPAID
GENERATING CODEBOOK TAXCB ON SAMPLE TAXPAYERS;
```

will produce a rectangular data file named TAXED containing one observation for each simulated family that paid some Federal tax in the last year of the simulation. Each observation will contain five data values corresponding to the five attributes listed in the CONDUCT command. A machine readable codebook file named TAXCB will also be produced; TAXCB will describe the sample survey file TAXED. These files may be used with other, independent computer programs on either the PDP-10 computer or another computer to perform any analysis for which programs exist.

Statistical and tabular outputs can also be generated directly within the MASH system. To compute a regression equation of personal wages as a function of age, race, sex and education for all persons working at least 47 weeks per year, it is only necessary to execute the following MASH statements:

```
DEFINE SAMPLE FULLYEAR OF PERSON AS WEEKS ≥ 47;

COMPUTE REGRESSION OF WAGES ON AGE. RACE, SEX,
HIGHGRADE ON SAMPLE FULLYEAR;
```

Suppose it is desired to tabulate the distribution of income by taxes paid for each family in the sample. Each attribute must first be coded, or classified into intervals. The intervals are defined in the form of a code:

```
DEFINE CODE MONEY AS (*-0=1  1-2000=2/  2001-4000=3/
4001-6000=4/  6001-10000=5/  10000-25000=6/  25000-*=7);
```

Each term of the code statement specifies a mapping, or functional transformation, of a range of money values into an integer value. For example, all values between 2001 and 4000 are to be mapped into the value 3. The symbol "*" represents either the lowest value possible or the highest value possible within the computer depending upon which side of the hyphen it appears. Codes are applied by defining new attributes as in the following examples:

```
DEFINE ATTRIBUTE CODEDY OF FAMILY AS TOTINCOME
CODED.BY MONEY;

DEFINE ATTRIBUTE CODEDTAX OF FAMILY AS TAXESPAID
CODED.BY MONEY;
```

In practice, a code having somewhat different intervals would be defined and applied to the tax variable. Generating cross-tabulation output is then performed by executing the command:

```
COMPUTE CROSSTAB  OF CODEDY, CODEDTAX;
```

if the cross-tabulation were desired for only those families which were "not in poverty" according to a standard definition, and if percentage distributions were desired, the following commands would obtain the output:

```
DEFINE SAMPLE NOTPOOR OF FAMILY AS TOTINCOME >=
1000 + 800*NUMPERSONS;

COMPUTE CROSSTAB OF CODEDY, CODEDTAX WITH
ROWPCTS, COLPCTS, CELLPCTS ON SAMPLE NOTPOOR;
```

For purposes of efficiency and automatic scheduling of output generation, output procedures can be grouped into censuses which can be scheduled to occur automatically. For example, suppose that the above regression and cross-tabulation are to be computed every two years during the course of a simulation. The MASH user would enter the following commands:

```
DEFINE PROCEDURE REG AS COMPUTE REGRESSION
OF WAGES ON AGE, RACE, SEX, HIGHGRADE ON
SAMPLE FULLYEAR;

DEFINE PROCEDURE XTAB AS COMPUTE CROSSTAB
OF CODEDY, CODEDTAX WITH ROWPCTS, COLPCTS,
CELLPCTS ON SAMPLE NOTPOOR;

DEFINE CENSUS BIENNIAL AS REG, XTAB;

TAKE CENSUS BIENNIAL EVERY 2 YEARS;
```

The MASH system is currently being extended to include an aggregate time series data bank which will form the data base for the macroeconomic portion of the model. Commands planned to access and manipulate the data bank include statements of the form:

USE DATABANK BELONGING TO ORCUTT:

ENTER SERIES GNP FROM 1929 TO 1937 AS 94.3, 91.7, 72.6, 78.5, 86.9, 88.2, 89.7, 96.1, 98.6:

CHANGE SERIES GNP IN 1931 TO 70.6:

RELABEL SERIES GNP TO GNPCONP:

TYPE TABLE FROM 1946 TO 1957 OF SERIES GNP, INVEST, CONSUMP, EXPORTS, MONEYSUPLY:

TYPE INDEX FOR MY DATABANK:

LAG GNP BY 1, CREATING GNP.LAGGED:

CLOSE DATABANK:

Little has been said about the integration of the operating characteristics of a microanalytic model into the MASH system. This step is accomplished in MASH without much difficulty by relying upon a traditional programming language, Fortran IV, and a custodial programmer to function as the interface between non-programming model builders and the computer programs containing the model. The derivation of operating characteristics for this class of models is sufficiently challenging and difficult that the model builder should not be restricted in his effort by being concerned with (and often restricted by) the details of the process of implementation.

Work is currently proceeding in several areas: (1) the refinement and extension of the set of operating characteristics basic to the structure of our model; (2) the integration of the auxiliary macro model with the microanalytic simulation model; and (3) programming and testing the MASH system which implements them. We expect to have an initial model implemented shortly. Following this first implementation will be a continuing process of extension, revision and modification as new useful knowledge becomes available and as new demands are placed upon the model by researchers and policy makers. It is our hope that those demands can be met successfully as a result of our present efforts.

The Urban Institute