APPENDIX 1. CONSTRUCTION OF THE MICROANALYTIC UNIT AND INITIAL POPULATIONS FOR SIMULATION

This appendix discusses the substantive and technical considerations underlying the structure of the simulation unit chosen for the microanalytic model described in [O5] and therefore contained within the MASH system. The methods used to create these simulation units for two initial micro populations are described.

Construction of the Simulation Unit

The choices of entity types, or units of simulation, are crucial choices for any simulation process. These choices substantially determine the potential complexity of the model's operating characteristics and the size and nature of the set of feasible simulated outputs. In general, increased complexity or disaggregation of the representation of simulation entity types allows a more detailed model specification, but it also places an additional burden upon the modeler to specify rules for changes of state for more entity types and their attributes. The choice of simulation entities and their attributes must also take into account the availability of empirical data with which to specify an initial simulation state.

The basic micro entity for the study of individual economic and demographic behavior is the *person.* However, a great deal of individual behavior depends upon the family relationships which a person has. For many purposes it is more realistic to regard groups of related individuals as decision making units rather than one individual in the group. Several entity types may therefore be useful in modelling and simulating household behavior.

Several conceptual decision making units within the household sector suggest themselves for consideration. For example, for demographic purposes the *nuclear family*, consisting of one or two parents with their own children, if any, is a useful construct. Other groupings of household members may also be relevant in studying the income generating, consumption, and saving behavior within the household.

Individuals obtain income from many sources. Among these sources are employment producing wages and salaries, businesses and farm proprietorships, capital, and a variety of transfer payments originating in the government, business, and household sectors. Most income received is money income, but some forms of income are "in kind," i.e. transfers of specific goods and services. Income may be obtained by an individual because of his or her characteristics, such as wages and dividend payments on owned stock, or it may be received because of the characteristics of other, often related, individuals, such as aid to families with dependent children and payments for foster child care.

Persons who receive income may not be those persons whose behavior generated the income. Salary payments received by an individual are payments for current services, but pension payments received by him or by his survivors after death may be in part deferred payments for his services. Dividends from stock received by a family may come from inherited assets. The unit that generates income may differ from the unit that receives it. Transfer payments from government may be paid to individuals because of their attributes, e.g. old age assistance, or they may be paid to survivors of the original beneficiaries.

The generation and distribution of assets is important in studying the distribution of income, since returns from assets are income. It may be useful to conceive of a household sector *asset accumulation unit;* such a unit would consist of an individual or group of individuals who own joint assets, make joint decisions regarding investment of current income in such assets or in assets owned by their dependents, who make decisions regarding asset transformations, and who determine the disposition of income from assets. Membership in asset accumulation units will change over time; such units may merge, split, or dissolve. The formation of new families will generally split existing units and create new ones, and various types of assets may be transferred to the new unit. Units are dissolved by the death of the last member of the unit, and the assets are transferred to other units.

The unit of consumption, or *spending unit,* is also important for studying the distribution of income. The spending unit concept is a construct of the University of Michigan Survey Research Center for the purpose of defining units of enumeration for the Surveys of Consumer Finances. Generally, the spending unit consists of all related members of a household residing together who pool their resources and make joint expenditure decisions. The spending unit is also likely to change over time as demographic and economic changes of state occur within the household. [1] The composition of the spending unit and its characteristics and preferences are important since its time preferences determine its saving rate and hence its acquisition of assets, and since the aggregate amount and composition of current consumption influences the general level of economic activity and therefore of personal income.

An important factor to consider in choosing the unit of simulation is the existence of data from which an initial simulation state may be derived. Household microdata are generally obtained from censuses and sample surveys performed by government agencies, private survey research institutions, and private individuals and firms. Within the United States, the largest and most comprehensive of these household data collection procedures are performed by the federal government. Because sources of current and

[1] For a discussion of the concept of the spending unit and some of its implications, see G. Katona, L.R. Klein, J.B. Lansing, and J.N. Morgan, *Contributions of Survey Methods to Economics,* Columbia University Press: New York, 1954, pp. 9-19.

comprehensive microdata are essential to the success of microanalytic models of this type, the structure of our unit for simulation is limited by the structure of the *enumeration unit* for which data are available from U.S. federal government economic and demographic household surveys.

During the past several years, three federal government household data collection procedures have dominated the set of possible initial state data for simulation of microanalytic models of the household. They are (1) the Decennial Censuses of Population and Housing [U2]; (2) the monthly Current Population Surveys [U1]; and (3) the 1966 and 1967 Surveys of Economic Opportunity [R1], [S2], [S4], [S5], [S6]. More recently, a number of "Public Use Samples" have been made available by the Bureau of the Census that contain specialized data subsets extracted from the 1970 Census of Population and Housing. These samples are expected to be quite useful for creating initial populations for microanalytic modelling.

Although there are minor differences in enumeration unit definition and structure among these surveys, they are basically similar in structure. The structure of the unit of simulation constructed for the model in [O5] reflects primarily household concepts and structure used in the Current Population Surveys and the Surveys of Economic Opportunity. This choice of structure conveniently allows these data sources to be used as initial simulation states without extensive manipulation of the data.

The basic Census enumeration unit for the above censuses and surveys is the *address.* Each address is classified as either a *housing unit* or not a housing unit (in Census terminology, an "other" unit). The current Census definition of a housing unit is:

> "A group of rooms or a single room is regarded as a housing unit when it is occupied as separate living quarters, that is, when the occupants do not live and eat with any other person in the structure, and when there is either (1) direct access from the outside or through a common hall, or (2) a kitchen or cooking equipment for the exclusive use of the occupants ..." [2]

A housing unit may be occupied by a family, by a person living alone, or by from 2 to 4 unrelated individuals who share the same living quarters. Any address in the sample found to be occupied by 5 or more unrelated individuals, such as a sorority house, is classified as an "other unit."

The current Census definitions of *household, family,* and *unrelated individual* are as follows:

> "Since 1960, a *household* includes all of the persons who occupy a house, an apartment, or other group of rooms, or a room, which consists of a housing unit under the 1960 census rules."

[2] U.S. Bureau of the Census, *Current Population Reports,* Series P-60, No. 60, "Income in 1967 of Persons in the United States," p. 4.

> "The term 'family' ... refers to a group of two or more persons related by blood, marriage, or adoption and residing together; all such persons are considered as members of the same family."
>
> "The term 'unrelated individual' ... refers to persons 14 years old or over (other than inmates of institutions) who are not living with any relatives. An unrelated individual may constitute a one person household by himself, or he may be part of a household including one or more other families or unrelated individuals, or he may reside in group quarters such as a rooming house."[3]

Every household that occupies a housing unit contains an individual who is designated the head of the household. By Census convention, no head of household is designated for households residing in "other units."

In order to retain the term *family* for describing a nuclear family-like entity, we use the Survey of Economic Opportunity term *interview unit* in its place. It is synonymous with the definition of *Census family* stated above.[4] Thus, an *interview unit* consists of two or more persons related by blood, marriage, or adoption who are residing together.[5] Census definitions provide for 4 types of interview units:

1. A *primary Census family* which includes the head of the household and all members of the household related to him or her.
2. A *primary individual* is a household head living alone or with non-relatives.
3. A *secondary Census family* includes a group of persons, related to each other but not related to the head of the household.
4. A *secondary individual* is a person who is not a household head and is not related to any other person in the household.

The sole exception to the interview unit definition given above is that secondary individuals under 14 years of age may be defined as members of the primary family. Such definition is at the discretion of the interviewer, and appears to occur when the secondary individual is a foster child. However, in the Survey of

[3] *Ibid.*

[4] In retrospect, the choice of the term *interview unit* to represent this level of simulation unit was unfortunate. The choice was made at a time when the 1967 Survey of Economic Opportunity was the best source of microdata from which to create an initial population for simulation. In adopting the 1967 SEO data file, some of the terminology describing it was unconsciously adopted also. While there is no operational difficulty in using SEO terminology to describe levels of the unit of simulation, some names used may be misleading because of the close similarity with standard Census names that have somewhat different meanings.

[5] *1967 Survey of Economic Opportunity Codebook,* Office of Economic Opportunity, p. 73.

Economic Opportunity, "if there are several persons under 14 years old, not related to the head, but related to each other," each of these persons remains a secondary individual.[6]

Each primary and secondary Census family has one of its members designated as the head of the family. The Census criterion for selecting the head is:

> "The head of a" (Census) "family is usually the person regarded as the head by members of the family. Women are not classified as heads if their husbands are resident members of the family at the time of the survey. Married couples related to the head of a family are included in the head's family and are not classified as separate families."[7]

Within a primary or secondary Census family, there may exist one or more *subfamilies:*

> "A subfamily is a married couple with or without children, or one parent with one or more single children under 18 years old, living in a household and related to, but not including, the head of the household or his wife. The most common example of a subfamily is the young married couple sharing the home of the husband's or wife's parents. Members of a subfamily are also members of a primary family."[8].

The Census definition of household structure allows *only* primary families to have subfamilies. Although it is not explicitly stated, subfamilies that would otherwise belong to secondary Census families must therefore be categorized as additional secondary families. Some connections between related subfamily members are therefore lost.

As an illustration of the structure of the full definition of Census household structure, figure A1-1 represents schematically the possible configurations that a household in the 1967 Survey of Economic Opportunity may assume. The sample counts are derived from codebook information and may vary slightly from actual file content due to changes made in the file subsequent to publishing the codebook and some remaining inconsistencies in the data.

If the complete Census household structure were to be retained as the unit of simulation, the simulation model would have to contain processes for forming and dissolving such units. For example, if the simulation process includes rules for dissolving households -- for example, due to marriage, migration, or death -- then it must also include rules for forming new households, or else the number and composition of households will become meaningless as the simulation progresses. Complexity in the

[6] *Ibid.*, pp. 73-74. Foster children are specifically identified by the interviewer, and a code designating this status is included in a household relationship attribute within the person segment.

[7] *Current Population Reports, op. cit.,* p. 6.

[8] U.S. Bureau of the Census, *Current Population Reports,* Series P-60, No. 66, "Income in 1968 of Families and Persons in the United States," U.S. Government Printing Office: Washington, D.C., 1969, pp. 6-7.

structure of the simulation units places a responsibility upon the operating characteristics of the microanalytic model to maintain the structure of those units in a realistic manner.

Figure A1-1. Structure of 1967 Survey of Economic Opportunity Household Unit

Because of this increased complexity and a lack of corresponding benefits, Census households have not been maintained but are initially divided into independent units of enumeration in deriving simulation units. These units each consist of one interview unit (or Census family) and the nuclear families and individuals within the unit. This decision has its basis in the assumption that non-related individuals or groups of individuals are far less likely to act as an interdependent group in generating or receiving income, accumulating assets, or spending than is a group of individuals related biologically or through marriage. Certainly exceptions to this rule in both directions can be cited, and such exceptions are occasionally observed in surveys:

> "There are also a few instances of unrelated persons who live together and pool their income as completely as members of any family. The 1951 Survey [of Consumer Finances] turned

> up, for example, two elderly women who lived together and pooled their pensions of $55 a month apiece for all expenditures of any sort. This type of arrangement, however, is rare.[9]

In recent years, communal living arrangements have become a more common example of such behavior.

The advantage of retaining a multi-interview unit household structure is that the structure of the simulated population then bears a closer resemblance to the structure of the real population of households in the United States. If the focus of the simulation were essentially demographic, then the benefits of such increased realism might justify the costs involved in constructing processes for household formation and dissolution. Since, however, the model's focus is economic, a decision was made not to maintain the household membership of interview units in the simulation unit.

Interview units (Census families), however, are not further subdivided for the purpose of forming units of simulation. The basis for this decision is that related individuals and families are quite often likely to act in an interdependent manner in their demographic and economic behavior. Thus, provision is made in the simulation unit structure for two (nested) groupings of individuals, the (nuclear) family and the interview unit. Microanalytic models defined upon such a population structure may contain operating characteristics that depend for inputs upon one or more related nuclear families and that cause an alteration in the structure of the families within the interview unit.

Because the focus of the model is essentially economic, it is of some importance to be able to account for both behavioral relationships and economic transactions between related persons not residing at the same address (and therefore not contained in the same interview unit). Although most surveys do not contain all relatives of all persons in the survey -- and those that implicitly do, such as the decennial census, do not attempt to ascertain any inter-household genealogical links -- such relationships are unlikely to exist in an initial simulation state. However, the history of demographic changes of state generated by the simulation process may be used to create genealogical links within the population. To the extent that these links exist, simulation processes can operate upon genealogically related individuals not residing together.

Structural alterations such as dropping the housing unit link between interview units in the simulation may already be partially accounted for by other characteristics of the simulation model. For example, if younger single persons living apart from their parents tend to "double up" in housing units, their expenditures on rent and related housing costs will generally be lower than if they occupied separate housing units. The distribution of housing costs for single individuals would therefore be lower at low

[9] Katona, et al., *Contributions of Survey Methods to Economics, op. cit.,* p. 18.

ages than it would be in the absence of group living, and a simulation process based upon the distribution would tend to replicate this downward shift in housing costs for some young persons. Although group living is not explicitly included in the model, its depressing effect upon housing costs would be implicitly incorporated into the simulation.

The other major change made in the Census enumeration unit to form the simulation unit relates to the subfamily. Separate balance sheets and income statements have been established at the subfamily level rather than being kept at the interview unit level. Furthermore, the definition of subfamily has been relaxed so that it may contain 1 or more members rather than 2 or more members. This revised subfamily concept corresponds to the concept of the nuclear family.

Subfamily members occupy an ambiguous position in the Census definition of family. Structurally, a subfamily resembles other primary families, but shares its living quarters with members of a primary family and it does not contain the head of the family; this implies some sort of dependence upon the primary family.

Young families represent potential family units that are either preparing to assume a geographically independent existence or that may have received a setback in attempting to do so. Sharing living quarters with a primary family can allow reduced expenditures and the accumulation of assets so that a transition to greater independence can be made later. To the extent that independent asset accumulation by young subfamilies is a significant factor in their financial behavior, including balance sheets and income statements at the simulation family level will allow such processes to be simulated.

Aged subfamilies generally represent a parent (or parents) or other relatives who have some physical, financial, or other dependence upon a member (or members) of the primary family. This dependence might be temporary or it might continue for the remainder of the subfamily's members' lives. Assets belonging to aged subfamilies might be kept separate from those of the primary family for several reasons: (1) to minimize taxation of income from assets; (2) as a reserve fund for large potential expenses such as medical costs; (3) to preserve the potential financial independence of the subfamily, should it wish to separate from its primary family in the future; or (4) because members of the subfamily might not want to relinquish ownership of their assets. These reasons imply that separate balance sheets for aged subfamilies will enhance the realism of the simulation process.

If separate asset, liability, and income data are maintained on a subfamily basis then the realism of the population structure is also enhanced by allowing subfamilies to contain one person. This allows individual children in primary families to become subfamilies as they begin to achieve financial separation and independence from their parents, while still living with them. Family formation can now be at most a three state process: (1) achieving partial or full financial independence from one's parents; (2) separating geographically from one's parents; and (3) uniting with another person through marriage.

Figure A1-2. Simulation Unit Structure

Figure A1-2 contains a schematic diagram of the structure of the new simulation unit. It consists of three levels, which for the purposes of describing the initial microanalytic model are called: (1) interview unit (Census family); (2) family (Census subfamily or nuclear family); and (3) person. It should be remembered that these specific names are associated with the microanalytic model and do not imply any restriction upon the more general applicability of the MASH simulation system. MASH can be applied to any microanalytic model in which the units of simulation can be cast in a three level hierarchical structure, regardless of either the names of those levels or the substantive entities which they represent.

Creation of Initial Populations for Simulation

The term *initial simulation state* refers to the collection of information required to specify completely the initial conditions for a simulation exercise. The initial simulation state is described by the microanalytic operating characteristics, the specification of the aggregate model, the values of the exogenous and initial (pre-simulation period) values for the aggregate endogenous variables, and the values of all attributes for all entities in the initial population. This section discusses the derivation of the sample survey files that were used to define the the last component of the initial simulation state -- the formation of initial populations from existing data sources. The description of the current microanalytic operating characteristics and the aggregate model appear in [O6].

Initial micro populations for simulation using MASH are formed by extracting micro data from an appropriately structured and documented survey data file and reorganizing them into a format suitable for efficient processing during simulation. The survey data file must meet several requirements: (1) the survey must be self-weighting, i.e. there must (implicitly) be equal sample weights associated with each person in the survey; (2) the survey file must be organized into segments of data corresponding to the three level hierarchical simulation unit described earlier and the segments must appear in left list order; and (3) the survey data file must be documented by a machine readable codebook structured according to the description in Appendix 3.

Since most survey data files produced are not structured precisely in the above form, they must first be reorganized to meet these requirements in order to be used as input to MASH. Such a procedure was first constructed in 1970 to reorganize the 1967 Survey of Economic Opportunity file into a form compatible with MASH. In 1972 another procedure was constructed to reorganize the 1960 and 1970 1-in-1,000 sample files of the Decennial Census of Population and Housing. Initial populations were extracted from these files and were included in initial simulation states for a variety of simulation experiments. Other more recently available micro data files such as the Current Population Survey and the Census Public Use Samples have a structure similar to the decennial census files and could be restructured in a form compatible with MASH with only moderate effort.

The following sections describe the procedure by which the 1967 Survey of Economic Opportunity file and the 1960 and 1970 1-in-1,000 sample census files were reorganized so that they could be used to generate initial populations for simulation.

The 1967 Survey of Economic Opportunity File

The 1966 and 1967 Surveys of Economic Opportunity (SEO) were conducted for the U. S. Office of Economic Opportunity by the Bureau of the Census in the spring of 1966 and 1967. These surveys include much of the information routinely collected in the annual February-March Current Population Surveys (CPS); in addition, they contain additional demographic and financial information not usually obtained between decennial census years. In both years, information was gathered regarding housing, assets and liabilities of interview units, and migration. In 1966, information was collected regarding job training; in 1967, data were collected on personal health, marriage, and child bearing.

The SEO units of enumeration were drawn from a random stratified cluster sample of U. S. households, as defined by the Bureau of the Census. In all, there are approximately 27,000 households in each year's sample. The sample consists of two independent subsamples: (1) the E-1 sample, consisting of approximately 21,000 households that are randomly selected from the CPS sampling frame to produce a sample approximately 50% of the size of the CPS sample; and (2) the E-2 sample, or "non-white supplement," consisting of approximately 14,000 households drawn from Census tracts containing a large percentage of non-whites in 1960. The discrepancy between the size of the total sample and its constituent parts is accounted for by a high level of non-response in both years.

Enumeration units in the SEO files are not uniformly weighted for a number of reasons. Stratification is introduced in three ways: (1) units in the E-1 and E-2 subsamples were selected with different rates of inclusion; (2) households selected in Primary Sampling Units (PSUs) that were represented in both E-1 and E-2 samples had their weights reduced to compensate for subsample overlap; and (3) in the E-2 sample, some PSUs called "self-representing" were included in the sample with certainty, while only a subset of the other, "non-self-representing," PSUs were included in the sampling frame. The base weights assigned on the basis of this sampling frame were then first adjusted for non-response and then normalized to provide aggregate population totals equal to those provided by independent population estimates based on age, color, and sex. Further adjustments to the weights were made to improve residence estimates and representation in the sample by family size.[10]

The weight associated with each SEO interview unit is by definition equal to the weight of the person who is the head of the interview unit. Prior to the adjustment of the person weights to match independent population estimates, the weights of all persons in an interview unit were identical. The subsequent adjustments produced small deviations in the person weights within each unit; inspection of the data suggest that most person weights do not differ from the weight associated with their interview unit by more than one-half percent. In order to simplify the selection procedure and because of the small deviations involved, the existing person weights were discarded and replaced with their corresponding interview unit weight.

One procedure for obtaining a uniformly weighted sample of SEO interview units (and therefore of persons also) is as follows. Let N be the number of interview units in the SEO file. Order the interview units in an arbitrary sequence, and denote the weights associated with these units as I(1), I(2), ... , I(N). Then

[10] More information regarding the SEO sample design, stratification, and weighting adjustment process is contained in [S4].

$$P = \sum_{i=1}^{n} I(k)$$

is the aggregate interview unit population estimate for 1967 from the 1967 SEO file. Let

$$C(j) = \frac{1}{P}\sum_{k=1}^{j} I(k)$$

The function C(j) is the cumulative distribution function of weights over interview units for that ordering of units. Note that C(0)=0, C(N)=1, and since all weights are positive, C is a monotonically increasing function. Therefore the inverse of the cumulative distribution function, denoted by V(u), maps the interval [0,1] into the integers 0, ..., N uniquely. If X is a random variable that is equal to 1 in the interval [0,1] and equal to 0 elsewhere, then for any interview unit J, the probability that V(X) maps into J is equal to the weight of J, I(J), divided by the SEO population estimate P, i.e. I(J)/P. Thus one method of selecting a uniformly weighted sample of M interview units from the population of N SEO interview units would be to choose M independent values X(1), ..., X(M) of the random variable X and then select the M interview units V(X(1)), ..., V(X(M)). The weight to be assigned to each of the interview units in this sample (and therefore to each person in each unit) is P/M.

This procedure is logically equivalent to drawing a random sample of size M from the original SEO interview unit population, with replacement. If M is small relative to N, the probability is high that no interview unit will be selected more than once. If M is a substantial fraction of N, then it is likely that some interview units having a high SEO weight will be selected more than once, and many interview units -- generally those with lower weights -- will not be selected for the sample. If M exceeds N, then replication of interview units with large SEO weights will occur more frequently, and replication of units with smaller weights will occur in some cases; some interview units will not be chosen, and these will generally be those with small SEO weights. The procedure will yield a sample of equally weighted units regardless of the relative sizes of M and N.

Although the above procedure yields a self weighting sample of SEO interview units, it is statistically inefficient in several ways. The method does not provide any control on important variables such as family type and composition. It does not attempt to minimize the variance due to the original SEO sampling frame being a cluster sample. Furthermore, some inefficiency is introduced by not minimizing the number of times a particular SEO interview unit is selected.

In order to increase the usefulness and efficiency of the samples drawn, the above procedure was modified in several ways. First, it was considered important to ensure that the composition of units in the sample corresponded closely to their composition in the entire SEO population. Units in the SEO file were therefore first ordered according to the following eight mutually exclusive family type strata:

1. Interview units headed by husband-wife families containing children under 14 years old.
2. Interview units headed by husband-wife families not containing children under 14 years old.
3. Interview units headed by female headed families containing children under 14 years old.
4. Interview units headed by female headed families not containing children under 14 years old.
5. Other male headed interview units containing children under 14 years old.
6. Other male headed interview units not containing children under 14 years old.
7. Male unrelated individuals.
8. Female unrelated individuals.

In addition, instead of choosing M independent values of the random variable X, M specific values were chosen. The sequence used was the regular sequence:

$$1/2M, 3/2M, 5/2M, \ldots , (2M-1)/2M$$

The effect of reordering the SEO interview units and selecting the above values from X was to ensure that the distribution of family types in the above strata -- an important consideration for the validity of the simulations -- matched the distribution of family types in the SEO file as accurately as the sample size permitted.

Within the above strata, the SEO interview units were also sorted on the geographic cluster number taken from the sampling frame, and within cluster, the units were sorted by age of their head. Combined with the above selection sequence, these additional sorts served to minimize the inefficiency of the sample due to the cluster design in the original SEO survey by choosing units uniformly over clusters. In addition, use of the above selection sequence served to minimize the number of times a replicated unit was included in a sample, a third source of inefficiency in the original method of selecting a self-weighting sample.

Using the above procedure, sample survey files of 100, 250, 500, 1000, 2500, 5000, and 10,000 interview units were initially selected from the 1967 SEO file. For each sample, distributions of commonly used variables such as age, race, sex, education, region, and family size were computed and compared with the corresponding distributions for the 1967 SEO file. The comparisons indicated substantial agreement between sample and population totals even for moderate values of M. In addition, the samples had the property that they were self-weighting, accurately reproduced the family composition distribution of their source, and minimized the statistical inefficiency of the original sample design due to cluster sampling.[11]

The actual technical procedure executed differed from the above description in one respect. Instead of sorting the entire interview unit record and all records associated with it, a three word extract record was initially created for each interview unit in the population. Using data compression techniques, the following attributes were extracted for each interview unit:

Identifying number of interview unit
Geographic segment number of unit
Structural type of unit
Number of persons in unit
Sample weight associated with unit
Age of head of unit
Race of head of unit
Sex of head of unit
Highest grade attended by head of unit
CPS area designator associated with unit
SMSA code for unit, if any

The file consisting of these 108 bit extract records was then sorted to yield the eight strata specified above. The interview units to be included within the sample were then chosen according to the above procedure, and the extract records associated with those units were copied into another file. This file was used to determine basic distributions and relationships that existed within the sample, to the extent that they could be derived from the extracted attributes. This file was then resorted into the original SEO file sort order, and the sorted file was then used to select the entire SEO interview units from the SEO file. This modification in procedure increased its efficiency substantially and allowed some preliminary analysis of the selected sample to be performed at low cost.

Structural changes were made to each interview unit selected from the SEO file for inclusion in the output sample survey file. The changes included: (1) discarding non-interview information; (2) discarding information for persons who had left their unit that year; (3) dividing multiple interview unit households into separate units of enumeration; (4) expanding subfamily structure; and (5) correcting structural data recorded erroneously in the original file.

Non-interviews constituted a large proportion of the original SEO sample. Of the 34,760 households in the 1967 SEO sample, 26,473 were successfully interviewed and 8,197 were not interviewed.

[11] If alternative random samples of the same size were desired, at least the following two independent alternatives exist: (1) any non-identity combination of stratum codes may be applied and the file resorted using this new order; and (2) the regular sequence above may be replaced by one starting with k*M, 0<k<1, with successive increments of M between sequential values.

Approximately 60 percent of the non-interviews resulted from the sample address being vacant at the time of the survey. The remaining 40 percent of the non-interviews resulted from one of the following reasons: (1) no one could be found at the address; (2) the occupants were temporarily absent; (3) an interview was refused; (4) the dwelling at the address had been demolished; or (5) other reasons. The overall non-interview rate was approximately the same in the E-1 and E-2 samples.

Using independent population estimates and information about interviewed households in the same cluster as the non-interviewed households, the base weights originally assigned to the SEO sample were adjusted to cumulate to aggregate U. S. population statistics in a manner designed to minimize bias caused by the existence of non-interviews. Although some location and housing information is available for non-interview addresses, no data are available for its occupants. Rather than attempt to complete such households by the creation of synthetic families and individuals, the non-interview data were discarded for the sample selection process, and the adjusted weights were used for drawing the self weighting sample.

Person records exist in the 1967 SEO file for all persons who were in the unit in 1966 but had left prior to the reinterview in 1967. Of the 92,857 persons recorded in the 1967 file, 3,445 had left and could not be reinterviewed. Although the information about such individuals might be helpful in constructing operating characteristics such as migration and family formation, only a limited number of characteristics of the person are known and no link exists in the sample to a record of the person's current characteristics -- nor would one be possible in general since households in the sample frame do not form a closed population. On the assumption that as many persons are likely to have joined units in the sample as are likely to have left, the records for individuals who have left the sample have been discarded in the self weighting sample.

Multiple interview unit households were divided into separate interview units for the purposes of self weighting sample selection and inclusion in the sample. The justification for this division is described earlier in this appendix. Of the 26,473 interviewed households in the 1967 SEO sample, 741 or less than 3 percent contained multiple interview units. The distribution of interview units per household is:

Interview Units per Household	Households
1	25,732
2	655
3	65
4	11
5	5
6	4
7	1
Total	26,473

The proportion of multiple households in the E-2 sample (non-white supplement) is about 50 percent higher than the proportion for the E-1 sample (the half-CPS sample).

For each interview unit, one or more *family* entities were created and placed in the unit hierarchy between the interview unit and the persons within it. If no subfamilies were present in the interview unit, all persons in the unit were assigned to the first and only family created. If one or more subfamilies were present in the interview unit, members of the primary family not in any subfamily were assigned to the first family created, and a new family was created to correspond to each original subfamily.

Balance sheet and income statement attributes that were defined at the interview unit level in the SEO file were transferred to the family level in the survey file to be used for creating initial populations. The values of all such attributes were transferred to the first family in each interview unit, and zero values were defined for these attributes for subsequent families in a unit, if any. This mechanical assignment procedure was used since it was not known at the time the extent to which the initial values of these attributes would be required as inputs to operation characteristics. Furthermore, a more appropriate assignment of these items to families would require substantial analysis of the characteristics of persons within the families created and an imputation of these items to families on that basis. Research in progress by Smith, Franklin and Orcutt [S8] promises to yield considerably more detailed and useful wealth characteristics of families which could be imputed to the micro population units during creation of initial simulation populations.

A number of other modifications were made to the SEO data at the time that the survey files were extracted. The "adult" segment of the SEO file was extended to apply to all persons, and default values were supplied for all attribute values for persons under 14 years old. A file of approximately 90,000 corrections that had been accumulated by the Office of Economic Opportunity was merged in during the extraction process. Certain family size corrections were made when the original data were found to be erroneous, Finally, some control information was added in order to provide better identification of the units created.

The technical procedure for accomplishing the extraction of survey file attributes and creation of survey files from the 1967 SEO file was complex but straightforward. A master mapping file was created and used to control both the extraction process and the creation of the machine readable codebook describing the extracted file. The master mapping file contained for each SEO attribute its name and the segment (hierarchical level) number into which it was to be transferred; attributes to be discarded were mapped into segment number zero. The extraction program used the mapping file to extract all information about the attribute from the SEO file's codebook, including its position in the SEO file. The attributes were then ordered according to their new segment numbers, field sizes were computed from the SEO codebook value descriptions, formats were created for the new file records, the machine readable codebook defining the new file was constructed, and attributes were extracted from each interview unit, restructured according to the specification in the mapping file, and written onto the new survey file. This three level tree-structured survey file was then read by MASH, using the file's machine readable codebook, to create a number of initial populations for simulation used in the current research effort.

The Decennial Census 1-in-10,000 Public Use Samples

Decennial censuses of population and housing are conducted by the U. S. Bureau of the Census in every calendar year divisible by 10. While the original mandate for the decennial census only required an enumeration of citizens and other persons by state, the scope of census taking activity increased steadily:

> "In 1820, questions on citizenship and industry were added to the population census schedule; in 1840, questions on education and disability; and in 1850, questions on marital status, place of birth, occupation and value of real estate owned. Supplemental questionnaires also came into use in 1850, and these grew in scope. In 1890, there were 11 different areas of "social statistics," with 14 housing inquiries and over 190 dealing with population items. By 1890, the basic population questionnaire itself was a document of over 30 items covering a wide range of topics."[12]

Decennial censuses of population and housing now form a rich source of social and economic microdata.

In 1962 the Bureau of the Census made available a 1-in-1,000 sample and a 1-in-10,000 sample of household and person records from the 1960 decennial census. In 1970 the Bureau announced that several 1-in-100 samples as well as the 1-in-1,000 and 1-in-10,000 samples would be made available, and that the 1960 samples would be reissued in a format compatible with the 1970 data. Now both the 1960 reformatted files and the 1970 samples are available.[13]

Both the 1960 and 1970 1-in-10,000 decennial census samples are attractive data sources for microanalytic simulation. They contain data at the appropriate level of disaggregation. The data are rich in demographic and economic detail. The 1960 sample may be used to generate outputs of the model for the period since 1960, thereby allowing some model validation to be performed, while the 1970 sample provides both a microdata comparison with the results of a 10 year simulation beginning in 1960 and a recent accurate source of microdata on which future predictions may be based. Furthermore, the commonality of format between years yields economies in conversion to satisfy MASH input requirements.

[12] U. S. Bureau of the Census, *1970 Census User's Guide;* U. S. Government Printing Office, Washington, D. C., 1970, p. 4.

[13] It has recently been announced that the Current Population Survey microdata files will also be made generally available for research purposes.

Finally the sample size of approximately 20,000 persons is manageable and allows further subsamples to be extracted at low cost.

Six different samples have been made available by the Bureau of the Census. Three samples were obtained from a 15% sample of households and three were obtained from a 5% sample; the samples contain somewhat different information from each other. In addition, each set of samples has a geographic dimension; one of the samples contains state of residence data, another contains neighborhood characteristics data, and the third contains county group and SMSA data. The 15% sample containing state data was judged to be most useful for deriving initial populations and was chosen for use. The conversion procedure described below was only applied to this sample, although it would have been applicable with only minor modifications to any of the six samples.

The conversion of the 1960 1-in-10,000 Decennial Census of Population sample to a survey file format suitable for MASH consisted of two steps: (1) creation of a machine readable codebook describing the resulting survey file format and content; and (2) creation of the file itself. No reweighting or enumeration unit selection was required since the 1-in-10,000 sample is a uniformly weighted random sample. Because use of the 1967 SEO file with the microanalytic model preceded the availability and use of the decennial census files, the attribute names and values and their associated meanings that were incorporated in the object model were largely derived from the 1967 SEO file description. It was therefore necessary to convert the 1-in-10,000 sample in such a way that the attributes and their value structures required for simulation matched those from the SEO file.

Attribute definitions and values were altered where necessary in one of two steps in the process: (1) during a file reformatting process performed by special purpose programming; or (2) during the initializing pass performed to create a new micropopulation. A machine readable codebook was first created for the target output file, using the 1-in-10,000 sample file description in [U2] and the 1967 SEO file codebook. The resulting 1-in-10,000 file's codebook contained three entity levels corresponding to the simulation unit levels of interview unit, family and person. Where the codebook's description of an attribute differed from that in the sample file documentation, the transformation was included in the file conversion program. Interview units and families within each Census household in the sample file were easily identified and separated with the help of the substantial demographic structural identification included with the sample file. As in the case of the SEO file, balance sheet and income statement information that could be identified with an interview unit was arbitrarily assigned to the first family in the unit, leaving the problems of allocating the items over families and persons to a later stage. Once the codebook and special program had been constructed, the survey file was produced from the 1960 1-in-10,000 sample.

A modified set of initializing programs was constructed to initialize any initial population derived from these survey files. These initializing programs are structured exactly like object model operating characteristics, and are inserted in the model flow control in such a way that they are invoked at the time of population creation. The file identification in the corresponding machine readable codebook is used to determine which set of initializing operating characteristics should be applied to the population being created.

The initializing characteristics include special programming for modifying some Census attributes to correspond to the SEO attribute structure used in large parts of the object model and for imputing values for new attributes required in the initial simulation state and not present in the Census files.