Session 11

USING ADMINISTRATIVE RECORDS FOR STATISTICAL PURPOSES

ADMINISTRATIVE RECORDS IN OFFICIAL STATISTICS --

HOW CAN OUR RESEARCH ON ADMINISTRATIVE RECORDS IN CENSUS 2000 LEAD TO GREATER USE IN 2010?

Presented by:

Ronald C. Prevost
Administrative Records Research Staff
Bureau of the Census

ADMINISTRATIVE RECORDS IN OFFICIAL STATISTICS --

HOW CAN OUR RESEARCH ON ADMINISTRATIVE RECORDS IN CENSUS 2000 LEAD TO GREATER USE IN 2010?

Presented by:

Ronald C. Prevost
Administrative Records Research Staff
Bureau of the Census

Abstract

The vision of the Administrative Records Staff is to develop a corporate administrative records infrastructure that will maximize the effectiveness of each survey, census, and estimation operation by 2010. This evolution requires immediate research support on the topics of administrative records, census and survey micro-data linkages, aggregate administrative records and survey data uses, content modeling, data processing and warehousing. Concurrent research efforts addressing data access and privacy issues relevant to use of administrative records in statistical programs are essential. This paper identifies the research required to address our current perception of the major technical questions for social and demographic statistics.

BACKGROUND:

The expanded use of administrative records to develop official statistics in the U.S. has been discussed for decades. The Census Bureau has limited its use of administrative records largely to aggregate data, such as vital statistics and elementary school enrollment, in the development of population estimates. In the past, Census Bureau use of individually identifiable records has been limited to:

- Demographic Programs -- the Population Estimates Program uses IRS income tax return information in the computation of implied net migration rates,

- Economic Programs -- the Economic Area uses IRS forms to constitute the universe of establishments and provide wage and salary information for a variety of purposes in its census and survey programs,

- Decennial Census Program -- the Decennial Census uses information from administrative records for institutions and other non-household sources.

Operations based on the use of administrative records have created highly effective, timely, and cost-efficient measurements for the past 50 years. Recent advances in automation and record linkage accuracy have significantly increased the potential cost benefits from expanded uses of administrative records micro-data in statistical programs. Additionally, census and survey recipients have become increasingly resistant to respond to data collection requests. The combination of these factors has increased the opportunity for significant benefits through using administrative records in measurement systems. The potential use of administrative records as a means to (1) reduce cost and respondent burden in the census and, (2) improve the consistency and quality of surveys and population estimates, has become extremely attractive in the current environment of reinventing and refocusing the way we do business. This possibility encourages us to seriously address content comparability and quality issues inherent in any use of administrative records.

The Plan for Census 2000, released in February 1996, includes a proposal to use administrative records as a data source for 5 percent of non-responding households through Integrated Coverage Measurement and to supplement item responses. This proposal is in addition to the use of administrative records from the U.S. Postal Service's Delivery Sequence File that will help the Census Bureau improve the list of residential addresses. This bold approach has led to heightened interest in research efforts, such as data linkage and database construction, as well as methodologies for incorporating administrative record information into new estimation applications.

The Census Bureau documented its initial experience with developing administrative records sources during the 1995 Census Test and is continuing research to determine how we can improve upon our approach. We have created the Administrative Records Research Staff to conduct research on the use of administrative records in demographic surveys, population estimates programs, and population and housing censuses. This approach is in concurrence with Recommendation 8.2 of the National Research Council (NRC) report, "Modernizing the U.S. Census," (see attachment). Based on our initial results and the many uncertainties associated with proposed changes in the Federal programs that generate many of the records, it appears that full use in Census 2000 may exceed current understanding and capabilities.

OUR VISION FOR 2010 AND BEYOND:

The vision of the Administrative Records Staff is to develop a corporate administrative records infrastructure that will maximize the effectiveness of each survey, census, and estimation operation by 2010.

We are in the initial stages of developing a plan to accomplish this vision. This evolution requires immediate research support on the topics of administrative records, census, and survey micro-data linkages, aggregate administrative records and survey data uses, content modeling, data processing and warehousing. Concurrent research efforts addressing data access and privacy issues relevant to the use of administrative records in statistical programs are essential.

The focus of our vision is on how the Census Bureau could reinvent and revolutionize data collection and processing operations using administrative records in preparation for the next millennium. As this system develops, we plan to blend the results from each survey, census, and estimation operation into a single, combined, final product that incorporates the best attributes of each individual measurement device. Ultimately in this type of system, the decennial census would become an extension of annual operations and would serve the purpose of benchmarking and quantifying the effectiveness of these operations on a national scale. This vision distinctly contrasts with our old view of a demographic system based primarily on the once-per-decade census, which was a "snapshot" of America.

This vision of a statistical measurement system is an aggressive approach designed to draw strength from integrating current survey or census operations with data from administrative record files. This vision was clearly endorsed by the National Academy of Sciences' recent reports on rethinking methods for decennial censuses, population estimates, and periodic surveys using administrative record sources. We anticipate this approach will:

- Reduce respondent burden by either eliminating individual contacts or reducing the information requested of individuals,

- Reduce the cost of data collection by using administrative record information in lieu of survey or census data,

- Increase data quality by employing individual measurement operations as evaluations of each other.

Our approach is driven by three premises:

1. Administrative records will have the greatest effect on Census 2000 through a variety of coverage improvement operations. Administrative records can be used to improve address lists, define sites that require special enumeration and marketing, target field follow-up operations, and enhance Integrated Coverage Measurement procedures. Data collection costs will be reduced as a result of these operations.

2. Administrative records have the potential to substitute for, or supplement data on, census and survey forms. The National Academy of Sciences has strongly recommended that the Census Bureau pursue increased use of administrative records for censuses as an alternative to direct enumeration. This recommendation was motivated by escalating costs of census enumeration, increased reluctance of the public to respond to census data requests, and budget uncertainties. This alternative methodology requires extensive development before it is ready for implementation in large-scale operations.

3. Administrative records in combination with direct measurement operations have the capability to create high quality, timely annual statistics of population, housing and their characteristics for census tracts and blocks.

We recognize that this reinvention process will require a substantial level of research to become a successful endeavor. We must also develop an approach that provides the necessary flexibility to circumvent inevitable roadblocks to our success. Therefore, we are approaching our vision from two stages--Research Toward 2000 and Toward 2010.

RESEARCH TOWARD 2000

A reinvention activity requires revisiting the premises underlying an operation, determining if current or proposed technologies can be implemented practically, and configuring ongoing operations in the most effective manner. Since measuring the resident population is one of our primary objectives, we will need to reexamine the current operations in light of expanded administrative record use.

Three options are available for statistically representing our nation's population, primary measurement, secondary measurement, or a combination. Primary measurements include operations that ask individuals specific questions through censuses and surveys. Secondary measurements employ data collected by other agencies such as administrative records, which the Census Bureau uses as a proxy for individual responses. Because administrative records are a secondary data source and are not collected for the purposes of enumerating the population and its characteristics, the responses provided by individuals on these records may not match the concepts measured through direct data collection. Using both primary and secondary record sources, we must design a system that provides the flexibility to produce a statistical representation of the United States under constraints of changing administrative records availability and funding for direct collection efforts. Initially, the best solution is the integration of primary and secondary measurements. Ultimately, the solution may rely more heavily on secondary data.

The research agenda must be structured to meet the short-term goal for research and evaluation in the context of the 2000 Census operation and the long-term goal of an administrative records database as a corporate resource for all statistical programs. The agenda should be developed to create a body of general knowledge applicable to our long-term goals. There are many questions and uncertainties about the process of creating population and housing databases from administrative records and their use by the Census Bureau's statistical programs to make counts and estimates. We have simultaneously begun research from two different approaches (1) micro-data modeling and (2) aggregate data modeling.

Micro-data Modeling

In the micro-data modeling approach we are exploring our ability to create population and housing databases with administrative records and to use these records as a proxy for direct measurements. (Note: numbers enclosed in parentheses relate each topic to the chief NRC Recommendation they address --see appendix.)

Issues identified for research:

- What is the geographic and population coverage, content, quality, and timeliness of information contained on Federal, state, local, and private vendor administrative records in relation to direct measurement instruments such as censuses and surveys? *(8.2)*

- Can we develop processing procedures that produce accurate matches of administrative records to the MAF and facilitate accurate geocoding? *(8.2)*

- Can we develop administrative record systems that provide accurate population and housing unit coverage at different levels of geography? *(8.2)*

- What matching variables such as name, address, date of birth, social security number, phone number, and record linkage procedures will provide accurate record matching and elimination of duplicate administrative records gleaned from multiple sources? *(5.6)*

- How can we best link telephone numbers to individuals at their residence or elsewhere? *(5.6)*

- How do we use administrative records to improve the sample design for various surveys (e.g., improved stratification of frame or identification of special interest populations)? (6.2)

- What is the magnitude of missing data items in administrative files? In the absence of key data items such as gender, race, and Hispanic origin, can procedures be developed for accurate imputation? *(7.1)*

- How do we use administrative records in estimation to enhance the quality of the estimates now made by demographic surveys or those anticipated from the American Community Survey? Are reductions in mean square error achieved? (6.2)

- How do we use administrative records in imputing for missing data from direct enumeration? (6.2)

- How do we construct households and families using administrative records? How do we estimate their characteristics using administrative records? *(8.1)*

- Do administrative records have the potential to meet emerging data needs of the type previously met through direct data collection? *(6.2)*

Benefits to Census 2000 -- An increased capacity to employ micro-data administrative records can be used to:

- Evaluate and improve the quality of MAF by enhancing the address list and geocoding addresses correctly, (8.2)

- Supplement Census 2000 operations for non-respondents, *(5.6)*

- Enhance implementation of Small Area Estimates Program statistics (see Aggregate Data Modeling section for discussion) in refining MAF evaluations and targeting arcas requiring field follow-up procedures, *(8.1)*

- Reduce the cost of collection operations by correctly attributing a telephone number to a non-responding address, *(5.6)*

311

- Aid in decennial census coverage evaluation (ICM) and non-response operations. *(5.6)*

Aggregate Data Modeling

In the aggregate data modeling approach, we are exploring a limited expansion of the Small Area Estimates Program. The Small Area Estimates Program of the Census Bureau is a cost-effective solution to providing annual statistics for Federal resource distribution. Expansion of this program will provide local governments with the capacity to enhance their decision-making process to further direct resources to areas of need. This is particularly important in an environment of devolution.

Issues identified for research:

- How do we expand the geographic detail of our subcounty estimates to develop accurate population and poverty estimates for census tracts and blocks? *(8.1, 8.2)*

- How do we develop estimates of housing, housing vacancy, tenure, and structural characteristics for counties, census tracts, and blocks? *(8.1, 8.2)*

- Can we improve the quality of the current product through an increased understanding of administrative records data concepts? *(8.1, 8.2)*

- Can we develop/enhance products that work in cooperation with ongoing surveys and the new American Community Survey to improve the quality of the surveys and develop feedback loops that also provide quality assessments of the small-area estimates? *(8.1, 8.2)*

- Can aggregate administrative records information be used as initial cost-effective coverage and content assessments for micro-data administrative records? *(8.1, 8.2)*

Benefits to Census 2000 -- Expansion of the Small Area Estimates Program provides products that can be used to:

- Evaluate the MAF along with the MAF Quality Improvement Program, *(5.5)*

- Preselect/target resources to areas requiring special enumeration procedures, *(5.3)*

- Target Census 2000 promotion activities, *(5.3)*

- Assist Demographic Analysis in enhancing survey-based ICM estimates, *(5.6)*

- Review the final statistics produced from Census 2000. *(7.1)*

TOWARD 2010:

We do not yet know whether a 2010 administrative records census is a realistic goal. What we learn between now and the year 2000 will drive the research agenda. If the planned uses of administrative records prove feasible, their implementation will result in a wealth of data for analysis. Immediately after the 2000 Census, the census records become a very valuable research record set to use in comparisons with administrative record files. Research questions will focus on the feasibility of an administrative records census in 2010. The cumulative knowledge acquired through the research agenda and the implementation process for Census 2000 will produce necessary evaluations to help us determine how aggressively we pursue the incorporation of administrative records into an ongoing statistical measurement system.

Appendix: Selected Recommendations from the National Research Council
Source: Summary: Modernizing the U.S. Census., 1995

Recommendation 5.3 The panel recommends that the Census Bureau incorporate successfully-tested procedures to increase the initial response rate in the 2000 census, including the use of respondent-friendly questionnaires and expanded efforts to publicize the mandatory nature of the census.

Recommendation 5.5 The panel recommends that the U.S. Postal Service and the Census Bureau continue to work together to improve the decennial census. We endorse the expanded role for the U.S. Postal Service in the 2000 census in several areas: (a) development, maintenance, and improvement of an accurate address file for the nation's residential housing units, (b) checking the address list prior to the census to improve accuracy, (c) delivery of the mailed forms, and (d) ascertainment of the vacancy status of housing units during the census.

Recommendation 5.6 We recommend that the Census Bureau undertake a thorough reexamination of the basic structure, organization, and process by which the decennial census is conducted to obtain the full cost-saving potential of the proposed redesigned census. As one part of its reexamination, the Census Bureau should develop a plan for the 2000 census that eliminates a substantial fraction of the $1.3 billion cost increase (in 1990 dollars) from 1970 to 1990 that is not accounted for by the growth in housing units and the decline in the mail response rate. The target for this plan should be much more than the $300 to $400 million we have already identified.

Recommendation 6.2 The panel recommends that the Census Bureau broaden its research on alternatives for more frequent small-area data to encompass a wider range than continuous measurement, as currently envisaged. In that context, the Census Bureau should examine the cost-effectiveness of alternatives, the ways in which they meet user needs, and the manner in which continuous measurement or other alternatives could be integrated into the nation's system of household surveys. The research program should be carried out in cooperation with the federal statistical agencies that sponsor household surveys and should include evaluation of the quality of important data elements, the frequency and modes of data collection, and the manner in which the results would be presented, as well as methods for introducing change over time.

Recommendation 7.1 The panel recommends that the Census Bureau expand its examination and testing of race and ethnicity questions to provide comprehensive information on: (1) public understanding of the concepts and acceptability of questions, (2) compatibility among the several census items and the utility of cross-tabulations, (3) comparability of census data to race and ethnicity data collected in other federal surveys or obtained from administrative records, and (4) the quality of data for small areas and specific groups. This research needs to

be given high priority so that the results may be incorporated into the review of Statistical Directive 15 currently being conducted by the Office of Management and Budget.

Recommendation 8.1 The panel recommends that the Census Bureau work to improve the amount, quantity, and frequency of small-area intercensal data:

- The Census Bureau should conduct experiments with federal administrative records for deriving more frequent small-area intercensal data estimates. At a minimum, the panel recommends that the Census Bureau geocode several large federal administrative record systems and use them to produce small-area estimates.

- The Census Bureau should work with state and local governments to enhance the quantity and frequency of small-area data.

Recommendation 8.2 The panel recommends that the Census Bureau give a single unit sole responsibility to exploit administrative records and produce small-area intercensal estimates on a frequent basis. Its work on administrative records should examine geographic consistency and quality. The unit should develop methods for increasing geographic content; establishing consistency of federal, state, and local administrative data; augmenting content on national records; augmenting usefulness of the resulting information through modeling; and computerizing approaches to database management to facilitate the use of administrative data in a census. If the content of administrative records can be improved for use in preparing small-area estimates, that is desirable, but the major purpose of the unit would be to produce small-area intercensal estimates.

Matched Data and Social Security Analysis: Nuts and Bolts

Bernard Wixon

(202) 282-7183
Bernard.J.Wixon@ssa.gov

Dedication: To Dorothy S. Projector, who brought the use of eligibility microsimulation models to the Social Security Administration and who taught me why we could not simulate social security and disability using surveys alone.

December 11, 1996

1. Introduction[1]

Much of the literature on data matching appeals to intended uses which seem narrow in scope. For example, matching enables analysts to refine income estimates by comparing estimates from two or more sources. Such an example suggests, however, that what is at stake is a seemingly marginal improvement in existing estimates. On the other hand, for analysts concerned with social security issues, discussions of matched administrative and survey data take on a tone of urgency. Without such data, analysts are largely unable to simulate effects of major structural changes in social security. Nor can they describe those who benefit most from the progressive structure of worker benefits, in terms of such basic characteristics as family income. Nor can those found disabled under SSA's medical-vocational criteria be studied in the light of recent household events which may have encouraged them to apply for benefits. Such analytical tasks represent basic public program analysis--estimating how benefits are targeted and why people apply. Analysts studying most public programs deem such tasks routine and address them, in large part, by using household survey data. Yet due to the data requirements of contributory programs and disability programs, such analyses cannot be undertaken for social security programs without matched administrative and survey data.

I have had first hand experience with these data gaps. Twenty years ago, I worked with a small staff developing a microsimulation model to study the distributional effects of a broad range of programs. These included means-tested programs (for example, Supplemental Security Income or SSI and Aid to Families with Dependent Children) and tax programs (specifically, the social security payroll tax, and the individual income tax). Notably absent from this list, however, were the country's largest cash benefit programs: the social security retirement program (in this case, Old Age and Survivors Insurance or OASI) and the two

---

disability programs (Disability Insurance or DI and the disability portion of SSI).

In this paper, I consider the long standing data gaps underlying these analytical limitations, as well as the potential contributions of exact matching of administrative and survey data. These gaps affect the most basic program analyses--such as evaluating program targeting and application incentives. Neither program data nor household survey data, taken alone, support a reasonable course of distributional and behavioral research. Program data lack a full set of attributes on well being, demographic traits, household composition, and recent household events. Household survey data include such attributes, but have only limited program information. Eligibility simulation models can be used to add programmatic information to survey data; however, this requires that the survey contain program-specific elements of eligibility and benefit determination. In comparing means-tested and contributory programs, I show that household surveys, without administrative data matching, can be used to simulate benefits for means-tested programs, but not for contributory programs. A separate problem unique to disability programs is also discussed: How can survey responses be used to estimate who would be found disabled under SSA criteria? One approach is outlined, involving matching of information on SSA determinations to household survey data. Taken together, the need to simulate social security benefits and to estimate the disabled serves as a rationale for a continued program of administrative and survey data matching.

Section two will discuss some of the basic ideas underlying the use of household data in program analysis, including brief sketches of distributional analysis, behavioral analysis, and eligibility modeling; readers with background in these areas may skip this section. In section three, I consider how household survey data, in combination with eligibility simulation modeling, support basic program analysis for means-tested programs. In section 4, I turn to the social security retirement (OASI) program to explain why--in the case of a contributory program--administrative data on earnings must be matched to survey data in order to undertake distributional or retirement modeling. Section 5 discusses a problem unique to the disability programs and outlines an approach taken by social security analysts to address it. Some concluding thoughts are presented in section 6.

2. Some Basics of Program Analysis Using Household Data

Why Data on Individuals?

Analysis using microanalytic data--data on individuals and households-- prevents the loss of information frequently associated with efforts to model program eligibility or the behavior of individuals with aggregated data. The relationships of interest stem from program provisions (e.g., eligibility for benefits) and behavioral responses (e.g., decisions to apply for benefits). Such relationships depend on *combinations* of traits of particular persons or families. Nonetheless, because disaggregated data are sometimes not available or because of the tight deadlines associated with policy discussions, estimates based on aggregated or group data are sometimes used. But when only highly aggregated data are available, information on combinations of traits or joint distributions is often lost and assumptions must play a greater role. The use of microanalytic data is intended to avoid this loss of information. Furthermore, analytic approaches that preserve the heterogeneity in the underlying data also serve the increasing interests of policy makers in understanding program effects on population subgroups. For this reason, when modeling decisions by individuals or program eligibility determinations, analysts often choose individual data--typically from household surveys--if such data are available.

Data Collection Mandates

Those who collect information in the course of program administration or through household surveys have different objectives and constraints. Typically, program administrators collect only the information needed to administer the program, that is, to determine eligibility and benefits.[2] Program data are limited both with respect to the universe and the attributes collected. On the one hand, information is collected only for those who apply for benefits (and, in the case of a contributory program, for taxpayers). On the other hand, the

---

[2] There are times when program administrators collect information "for statistical purposes." However, under resource pressure, collection of some statistical information suffers, since it has no direct bearing on the determination of eligibility and benefits. In many cases, collection of such data degrades over time and the data become unusable.

information collected is mainly limited to what is needed to determine eligibility for benefits and the size of the benefit.

Household surveys, in contrast, collect information on a wider range of attributes. Information on such attributes is needed to investigate (1) distributional effects of programs on subgroups of interest (for example, race, income classes, education, marital status) and (2) incentives to apply for program benefits as an alternative to work and to consider other behavioral effects. And, although surveys are limited by the use of samples, the samples often represent the general population. However, surveys lack detailed program information, such as whether a respondent is eligible or how the respondent's benefit would be affected by a specific program provision. Indeed, while some such program information is uncollectable using standard household surveys, it is nonetheless central to program analysis.

There are other important differences as well. Because of recall constraints, surveys generally collect data relating to the time of collection. Survey cost constraints preclude data collection efforts involving both a large sample and repeated surveys over many years. This is possible in the course of public program administration, however, if such data collection is integral to program administration.

The Microanalytic Toolkit

Distributional analysis and behavioral analysis represent some of the most basic approaches to evaluating the effects of public programs or of proposals to change such programs. As they are now practiced, however, both approaches are often linked to eligibility modeling. Below is a brief sketch of each. It is not my purpose to provide a detailed description, since this paper considers the data requirements of these program analytic tools, rather than the tools themselves. However, I do mean to suggest that distributional analysis and behavioral analysis represent the most basic means by which program analysts estimate whether a public program is targeted to those intended, as well as the behavioral incentives created by the program.

(1) Distributional analysis.--Distributional analysis involves demonstrating how eligibility or benefits from a program or a specific program provision affect population subgroups of interest. In one sense, it is not obvious why analysts would examine such effects, since eligibility and benefits are carefully determined based on specific traits, such as advanced age and low income. That is, if benefits are carefully targeted to the low income aged, for example, why should analysts or policy makers be concerned about effects on other subgroups?

Broadly stated, distributional analysis addresses the issue of target efficiency of a public program or of specific program provisions. Target efficiency involves two elements: Ensuring that all intended recipients receive benefits and that *only* intended recipients receive benefits. But implicitly, the "intended" population is defined multidimensionally--not just in terms of explicit eligibility criteria, but in terms of the broad interests of policy makers.[3] For that reason, program data have limited value for distributional analysis. The point of distributional analysis, then, is to evaluate *government decisions* with respect to eligibility and benefits from the standpoint of broader criteria than those explicitly used in determining eligibility and benefits.[4] To support such analysis, the data source must include a comprehensive set of variables on well being and demographic traits and the universe should include nonparticipants to serve as a comparison group. Eligibility simulation models are

---

[3] For example, how the taking of early, reduced retirement benefits might affect poverty rates might be of considerable interest. Yet, because program data do not include measures of income from all sources, it is difficult to investigate such issues using program data alone.

[4] Because distributional analysis often consists of tabulated characteristics of persons falling into programmatic or socioeconomic categories, it is sometimes called "descriptive" analysis. This characterization is misleading in the sense that, as policy analysts know well, normative conclusions are often drawn from such information. Logically, this is possible because policy makers or analysts bring to the table well established views of the purposes of the programs. A logician would call such views latent normative premises. Let us consider an example.

•policy maker's latent normative premise:	a program for the low income aged should treat singles and couples equivalently.
•estimates show:	a much higher percentage of single recipients are in poverty
•policy maker's normative conclusion:	there is reason to question the design of the program with respect to family composition.

often used to define those affected by a specific program provision.

(2) Behavioral Analysis.--Behavioral analysis considers decisions of individuals, such as whether to apply for benefits and to quit or reduce work.[5] Such decisions are unlike government decisions, for which the determining factors are fixed largely by statute and regulation. The decisions of potential applicants to apply for program benefits are private, psychic events.[6] For that reason, the analyst uses statistical techniques to infer both the major factors involved in the decision and the roles played by such factors.

Optimally, the data should include information on the major factors hypothesized to determine application decisions and the data should distinguish applicants from nonapplicants. Such factors typically include measures of economic resources, demographic characteristics, recent household events, and the size of the potential program benefit. In the context of applications analysis, the size of the potential benefit has consistently been shown to be a major determinant of the decision. For that reason, eligibility models are frequently used in the analysis of decisions to apply for benefits. One implication of this is that behavioral analysis will be limited, unless the data set includes the information needed to estimate eligibility and benefits.

Once a model has been estimated, the decision factors, their relative importance, and the direction of their effects are known. The model can then be used as a policy simulation tool. In the course of simulating the short term response to a policy change, for example, the model might be used to estimate how many additional persons would apply if benefits were increased. Long term applications, in contrast, involve projecting the changes in applications that are implied by expected changes in the decision factors.

---

[5] While behavioral analysis can focus on savings behavior, family formation, and other decisions, I will mainly discuss the decision to apply for benefits.

[6] Many now believe that some state welfare organizations, hospitals, and firms mandate or encourage application for federal benefits. To the extent this is true, the application decision has become institutionalized, to some degree.

(3) Eligibility Models.--In most programs, eligibility determination and the computation of benefits is a lengthy and seemingly arcane process. However, by focussing on program detail in the foreground, it is too easy to lose perspective: program eligibility criteria reflect society's normative decision making--they determine who receives benefits from public programs. Hence, eligibility provisions represent choice variables through which policy makers can refine distributional effects or cut costs, for example. Eligibility determination has two parts, categorical and financial.

Categorical (or nonfinancial) criteria embody an underlying premise--that the programs are not intended for all, but for those not able or not expected to work.[7] Hence the categorical criteria for public programs correspond to population subgroups broadly considered to be dependent on support from other members of society or entitled to social insurance benefits. The criteria for some major programs are old age, disability, and single parent status.

Financial eligibility and benefits can be estimated using a simulation model if the data set has the needed elements. Financial eligibility and benefits for means-tested programs, for example, are determined by such factors as low income, low assets, family size, and living arrangements. In government decision making of this type, the decision process is detailed in legislation and regulations and should be applied uniformly for all applicants. Assuming the data requirements are met, the decisions are largely replicable using an accounting framework that mimics the eligibility and benefit algorithms.

These analytical tools represent one type of microsimulation model, since they simulate program eligibility using information on persons or families. From a historical perspective, such microsimulation models have provided a framework for incorporating data on individual persons and households into the distributional, behavioral, and cost analyses of benefits and taxes. These models use information on individuals and families on a case-by-case basis, taking each case through an eligibility and benefit determination procedure which mimics the

---

[7] Some programs, such as food stamps and the individual income tax, do not have such categorical criteria. However, most benefit programs, including those discussed here, are categorical.

claims process itself.

Survey-based microsimulation models give policy makers an effective means of considering hypotheticals, such as:

- *the program eligibility of individuals who have not applied for benefits.*--Eligibility models are used in conjunction with surveys representing the general population, allowing analysts to estimate the number of eligibles and their hypothetical benefits. Such estimates are useful for evaluating potential program growth. Also, eligibility estimates allow construction of participation rates, allowing analysts to evaluate a means-tested program in terms of how well it reaches its targeted population.
- *the eligibility, benefits, behavioral responses, and costs associated with benefit structures which have been proposed, but not enacted.*--An eligibility model's step by step representation of the eligibility and benefit determination procedures allows analysts to consider "what if" alternatives to the current structure. In this way analysts can either consider broad structural alternatives or incremental changes to specific provisions of the benefit structure.

3. Use of Household Surveys to Analyze Means-Tested Programs

In this section, I will discuss the data requirements for analysis of a public program, focussing on attributes of the sort collected in household surveys, as well as methods for introducing detailed information on eligibility and benefits. I will consider the SSI program (aged portion) which serves as an example of means-tested programs in general. This discussion will show why household data, without administrative data matching, can support an extensive course of research for some public programs. Using the SSI/aged case as a baseline, I will then consider (in the next section) programs with more vexing data requirements.

Attributes

First, it is helpful to distinguish attributes collected from respondents to household surveys

and those collected from program applicants in the course of program administration. But rather than classify a large number of variables, I would like to discuss broad groups of variables, mainly in terms of their roles in the analysis of public programs. Let us consider the two tiers of variables displayed in Table 1. The top tier includes examples--though not an exhaustive listing--of variables which can be used to determine program eligibility and benefits for SSI (aged portion). Categorical eligibility requires information on age, while financial eligibility is determined on the basis of income, assets, family composition, home ownership, and other factors. Such information is, of course, collected from applicants by program administrators. However, many household surveys collect comparable information.[8] This has a marked impact on the scope of program analysis which can be pursued for means-tested programs.

The second tier of variables, labeled "evaluative," are used differently. From the standpoint of their source, they may be thought of as variables collected by a household interview survey designed to support program analysis. From the standpoint of their use, they are likely to be used by analysts in two ways: (1) to define subgroups used in distributional analysis (for example, age, education, marital status, income class) and (2) as explanatory variables in behavioral analysis (for example, income, job loss, loss of health insurance). That is, these are the variables needed to support distributional or behavioral analysis for any public program. However, since most of these variables play no role in determining eligibility or benefits, they will not be available in program data.

Populations

Public program analysis also entails defining relevant subgroups. For example, program data include information on participants and, in some cases, limited data on denied applicants. General population surveys, taken alone, provide only a distinction between program

---

[8] The information collected in surveys is typically not identical to that collected for program administration and surveys often do not collect items with a minor role in determining eligibility and benefits. The Survey of Income and Program Participation (SIPP) was explicitly designed to collect the elements of eligibility and benefit computation for a number of programs.

Table 1

Variables Classified in Terms

of their Analytic Use

Attributes

Program Eligibility Variables

Categorical Eligibility
  (age)
Financial Eligibility
  (income, assets, family size
  home ownership . . .)

Evaluative Variables

Well Being
  (income, assets . . .)
Demographic Traits
  (age, race, ethnicity,
  education, health . . .)
Household Composition, Events
  (marital status, loss of spouse,
  job loss, loss of health ins. . . .)

beneficiaries and nonbeneficiaries and a measure of the size of the benefit. But when surveys include variables needed to support determination of eligibility and benefits through the use of eligibility models, those eligible--irrespective of whether they receive benefits--can also be estimated. Table 2 suggests the key populations.

Table 2 illustrates how questions in a household survey, combined with an eligibility model, can be used to define these populations. Whether respondents participate or not is based on the response to the survey question on benefit receipt. On the other hand, respondents not receiving benefits may not know whether they are eligible for a program. Program eligibility, then, is an attribute that often cannot be collected in surveys; it must be modeled. As Table 2 illustrates, the survey question on benefits, in conjunction with eligibility modeling, allows analysts to define four populations:

- Eligible participants.--This group includes the great majority of beneficiaries.
- Eligible nonparticipants.--The size and traits of this group are of interest, first, because they suggest the potential for program growth. Second, they represent a key control group vis a vis participants; that is, differences between eligible nonparticipants and participants are important to understand incentives created by the program.
- Ineligible nonparticipants.--This group, which includes most members of the general population, is useful mainly as a control group. The boundary between ineligible nonparticipants and eligible nonparticipants varies depending on the eligibility criteria simulated. Hence, analysis of alternative definitions of eligibility can only be undertaken if nonparticipants are included in the sample.
- Ineligible participants.--Participants can be found ineligible for several reasons including reporting errors, survey/program inconsistencies (if the survey reference period differs from the program accounting period, for example), errors in the modeling of eligibility, or fraud. This is typically the smallest of the four groups and our data often do not allow valid inferences about this group; for that reason, they will be excluded from further consideration here.

Those affected by specific program provisions.--Increasingly, program analysts study not just those receiving benefits from a given program, but those affected by specific provisions of the

Table 2

Populations Relevant for Program Analysis

		Is Person Participating?	
		yes	no
Is Person Eligible?	yes	Eligible Participants	Elig. Nonparticipants
	no	Inelig. Participants	Inelig. Nonparticipants

program. That is, we not only look at subgroups defined in terms of, for example, demographic traits or family income--we also look at subgroups affected by provisions such as an asset test or marginal tax rate, to consider distributional and incentive effects of such provisions. This approach is "analytical" in the formal sense of that word, breaking a whole-- in this case, the program's benefit structure--into parts.

SSI (Aged)

While the principal focus of this paper is to explain why matched survey data are critical for analysis of social security programs, for purposes of comparison I will discuss how, in the case of means-tested programs, household survey data, unmatched, support a range of program analyses. I will use Table 3, which combines the populations and attributes discussed above, to consider how well program data and household survey data, each used independently, satisfy the data requirements for distributional analysis, behavioral analysis, and eligibility modeling. The fundamental requirement involves having a comprehensive set of evaluative attributes for population subgroups defined in terms of fairly detailed programmatic criteria. That is, information on benefit receipt is not sufficient for purposes of program analysis. One approach is to link evaluative information available in surveys to detailed programmatic information for the same individuals.

On the face of it, data linkage might seem an obvious solution. However, data linkage (in this case, exact matching) is problematic for two reasons. First, it is useful only for those who have applied for benefits; it would tell us nothing about the program eligibility of those who have not applied. Second, matching is useful for some analyses of the existing program, but would not allow analysis of many proposed alternatives. Eligibility simulation offers a more flexible approach. Because it involves a representation of the eligibility process itself, eligibility models can be used to determine eligibility or incentive effects for those who have not applied or the effects of benefit structures which have not been enacted. For such analyses, data--whether matched or unmatched--must be supplemented with models that estimate eligibility and benefits.

Table 3 allows us to reflect on the data requirements for analysis of SSI (aged) by considering program data and household survey data in turn. Each can be considered in terms of attributes and populations. Program data, represented by the area outlined with the broken line, include only participants and only the attributes needed to determine categorical and financial eligibility.[9] However, both distributional analysis and behavioral analysis require data that: (1) juxtapose program information with evaluative variables for the same individuals and (2) include nonparticipants (as a control group for behavioral studies or for simulating eligibility alternatives). These factors limit the analytic potential of program data when used alone, although such data remain useful for describing program trends and predicting costs.

On the other hand, household survey data, unmatched, have considerable potential for analysis of means-tested programs. By virtue of the mandate under which much survey data are collected, they include evaluative variables basic to both distributional and behavioral analysis and they collect data on nonparticipants as well as participants. Taken alone, however, the program information included is too limited; it does not indicate, for example, how participants' benefits were affected by specific program provisions. Fortunately, in the case of means-tested programs, such program details can be simulated using eligibility models. The large shaded area in table 3 illustrates, then, that eligibility models can be used to define key programmatic populations and that the survey offers comprehensive evaluative variables. As a result, an unmatched household survey, such as the SIPP, can support a broad range of program analyses--including distributional analysis, behavioral analysis, and eligibility modeling--for most means-tested programs.

It is important to understand why means-tested programs can be simulated using household surveys. Both household surveys and means-tested programs are instruments of public policy

---

[9] In table 3, as well as in other tables used below, I discuss the analytical potential of data sets, used in conjunction with models, by illustrating to what extent the data set "fills" the attribute/population space shown on the table. In these tables, I show that the most useful data sets (1) allow estimates for all program-related populations shown and, (2) for members of each such population, include not only detailed program eligibility variables, but also a comprehensive set of evaluative variables.

Table 3

Supplemental Security Income (Aged):

Household Survey Data with Benefit Simulation

<u>Populations</u>

<u>Attributes</u>	Program Participants	Nonparticipants	
		Eligible	Ineligible
Program Eligibility Variables			
Categorical Eligibility (age)	**Program Data**	(Eligibility Simulated)	(Eligibility Simulated)
Financial Eligibility (income, assets, family size home ownership . . .)			
Evaluative Variables			
Well Being (income, assets . . .)			
Demographic Traits (age, race, ethnicity, education, health . . .)	**Household Survey Data (with Benefit Simulation)**		
Household Composition, Events (marital status, loss of spouse, job loss, loss of health ins. . . .)			

and, in some respects, they are focussed similarly. Household surveys collect information on the financial well-being of demographic groups, among other things. Means-tested programs are typically targeted to specific demographic groups with low income and low assets. And, both the survey and the program relate to the *current period* (typically, in a well designed survey, the survey reference period subsumes the program accounting period). These coinciding features allow the program simulation.

4. Special Case One: Social Security and Matched Survey Data

In section 3, I suggested that the data requirements for basic microdata analysis of means-tested programs are substantially met by household survey data if those data are used in conjunction with an eligibility simulation model. That discussion also serves as a benchmark for considering why household survey data, taken alone, do not support basic program analysis for the nation's largest cash benefit program, social security.[10] But before considering this in more detail, let me discuss the general criteria underlying the social security benefit structure, that is, how the program is targeted.

Dual Targeting of Social Security Benefits

      The report of the 1979 Advisory Council on Social Security included a statement of the principles underlying the OASDI benefit structure:

> From its beginning those responsible for the design of social security have sought to assure, on the one hand, a reasonable relationship between the social security taxes paid by individuals and the benefits they receive and, on the other hand, at least a minimally adequate income for long-term low-wage workers. Maintaining a reasonable relationship between taxes and benefits has been described as the goal of individual equity. Assuring a basic level of

---

[10] The discussion in this section relates to the social security retirement (OASI) program and to financial eligibility for the disability insurance program. A separate problem unique to the major disability programs, DI and SSI, is discussed in the next section.

income has been called the goal of adequacy.[11]

While covered earnings (and resulting contributions) represent the major determinant of benefit amount, a number of provisions are motivated by the adequacy goal. Major adequacy provisions include the "progressive" structure of worker benefits and dependent benefits. Other smaller-scale provisions include the windfall benefit provision limiting benefits for some federal workers, the minimum benefit (now repealed), and a special benefit for those over the age of 85 (occasionally proposed). Because of the scale of the OASDI program, some adequacy-related provisions have major effects. For example, over twelve million persons receive dependent benefits as spouses, children, or widows.[12]

Adequacy provisions under the structure of social security benefits are not explicitly targeted so as to take into account family income, as are provisions of a means-tested program. Some provisions are designed to provide additional benefits to those with low levels of covered wages or low benefits. Examples include progressive worker benefits, the repealed minimum benefit, and the proposed benefit for those over age 85. Analogously, family composition provides the targeting mechanism for dependent benefits. For adequacy provisions, then, low wages, low benefits, and family composition can be considered "proxies" for low levels of financial resources. And if they serve as proxies, then such provisions are target efficient to the extent that a broad, stable relationship exists between, for example, low covered wages and low levels of all financial resources. However, the denial of windfall benefits to insured federal workers serves as a reminder that targeting mechanisms such as low wages should be subjected to scrutiny. Matching administrative and survey data allows such scrutiny.

Let us consider Table 4. In terms of the variables needed, financial eligibility and benefits are determined differently for a contributory social insurance program than for the means tested program considered in section 3. Financial eligibility and benefits are determined based on such factors as the period working in a job covered under social security, earnings

---

[11] See page 55, Advisory Council on Social Security (1979).

[12] See Table 5.A1, Social Security Administration (1995).

Table 4

Social Security Benefits:

Program Data versus Household Survey Data

<u>Populations</u>

<u>Attributes</u>	Program Participants	Nonparticipants

Program Eligibility Variables

Categorical Eligibility
  (age)

Financial Eligibility
  (earnings history,quarters covered,
  family comp., marital history . . .)

Program Data

Evaluative Variables

Well Being
  (income, assets . . .)

Demographic Traits
  (age, race, ethnicity,
  education, health . . .)

Household Events, Composition
  (marital status, loss of spouse,
  job loss, loss of health ins. . . .)

Household
Survey
Data

histories, marital history, age at retirement, and current family composition. And, in terms of evaluative variables, in order to deal with dual targeting, analysts require information on *both*: (1) family income and other financial resources as well as (2) past earnings and contributions.[13] In addition to an estimate of potential benefits, behavioral analysis--for example, retirement modeling--requires information on variables such as household composition and recent household events.

Table 4 illustrates why, in the case of the social security program, *neither* program data nor household surveys, taken alone, support some of the most basic distributional and behavioral analyses. As suggested by the smaller block, program data include detail on benefits for those receiving benefits; however, they lack the evaluative variables needed to undertake distributional or behavioral analysis. Social security program data do not include information on total family income, assets, or recent household events, for example.

Household surveys, however, cannot support social security analysis as they can for means-tested programs. The impediment is that household surveys do not collect the information needed to simulate financial eligibility and benefits for social security, namely, earnings histories and the period working in a job covered by social security.[14] This limitation affects social security analysis in fundamental ways. Retirement modeling requiring an estimate of the retirement benefit for all eligibles cannot be undertaken. In terms of distributional analysis, characteristics can be tabulated for participants and nonparticipants, but not for more detailed beneficiary groups, such as workers with low lifetime wages or those receiving only

---

[13] Earnings, then have more than one role in evaluating a contributory program. On the one hand, earnings histories allow simulation of a social security benefit. On the other hand, analysts and policy makers will want to consider how specific provisions or proposals affect long term, low wage workers, as opposed to high wage workers.

[14] While some of these items could be collected in household surveys, it seems problematic to collect information such as the period working under social security or earnings histories, due to recall constraints. This also seems wasteful, in view of the fact that such information has already been collected administratively.

spouse benefits. Nor can analysts simulate alternative benefit structures.[15]

Table 5 illustrates what many analysts of social security take to be the solution--matching program data on earnings to household survey data. Assuming the survey has information on family composition and marital history, benefits can then be simulated for survey sample members. As the large shaded area illustrates, matching earnings histories and simulating benefits fulfill the critical data requirement: On the one hand we can simulate eligibility and benefits for sample members and, on the other hand, the survey provides evaluative attributes.

5. Special Case Two: SSA's Disability Programs

The Problem of Categorical Eligibility

Eligibility for major public programs involves two types of criteria, *categorical* and *financial*. As discussed above, for the social security retirement (OASI) program the obstacle to distributional and behavioral analysis involves the inability to simulate *financial* eligibility and benefits without using matched data. The problem relating to financial eligibility also exists for one of the disability programs--Disability Insurance--because the program is financed through contributions from past earnings. However, in this section I consider *categorical* eligibility, which poses a problem unique to the disability programs, DI and SSI (disabled). This problem, as addressed in recent SSA research, has involved the use of matched data other than information on earnings histories.

---

[15] A small number of program provisions can be simulated using survey information alone, such as how changes in cost of living adjustments might affect poverty. The key is that in a few such cases, simulated benefits can be estimated without information on lifetime earnings, by using benefit information reported in the surveys. For other issues requiring information on lifetime earnings, such information can be added through statistical matching or by imputing earnings streams. For most purposes, however, exact matching of the observed earnings is clearly preferred, especially when distributional analysis is needed.

Table 5

Social Security Benefits:

Matched Survey Data with Benefit Simulation

Populations

Attributes	Program Participants	Nonparticipants	
		Eligible	Ineligible

Program Eligibility Variables

Categorical Eligibility
(age)
Financial Eligibility
(earnings history,quarters covered,
family comp., marital history . . .)

Evaluative Variables

Well Being
(income, assets . . .)
Contributions
(earnings history, taxes)
Demographic Traits
(age, race, ethnicity,
education, health . . .)
Household Events, Composition
(marital status, loss of spouse,
job loss, loss of health ins. . . .)

Program Data | (Eligibility Simulated) | (Eligibility Simulated)

Matched Survey Data (with Benefit Simulation)

Disability determination, which determines categorical eligibility, serves a gatekeeping function for both the DI and SSI programs, distinguishing allowances from denials. With the total number of applicants for the two programs now having reached 2.5 million per year, its budgetary and income distributional effects are undeniable. However, using survey information to estimate those categorically eligible--those the Social Security Administration would consider disabled--has proven especially problematic with respect to disability programs. Other programs--those for the aged or for single parents, for example—use criteria relating to easily observed traits and, as a consequence, survey responses on such traits permit reasonable estimates of those categorically eligible.

But, how to use survey responses on health and activity limitations to represent program disability criteria is by no means self evident. Judgmental factors are involved in two respects. First, the critical survey responses are judgmental. When surveys ask respondents about their health and the extent to which impairments limit activities, the responses are self-evaluative or self-rated. Under one hypothesis, for example, of those with a given impairment, persons with a low tolerance for pain or a weak attachment to the labor force report more severe limitations. Second, the criteria used by SSA to determine disability status are complex and also involve judgmental elements. For example, early in the complex determination process, applicants are denied if their impairments are considered nonsevere. Also, at a later step, there is an evaluation to determine whether the applicant has the residual capacity to perform substantial work, after taking into account the applicant's impairment, age, education, and past work. Hence, both the program criteria and the responses to health questions in household surveys involve judgmental components.

It is not surprising, then, that survey estimates of the size of the disabled population cover a wide range. Table 6, which summarizes estimates from a recent Bureau of the Census report, illustrates this point.[16] The five estimates, each using conventional definitions, cover a disturbingly large range—from 3.4 million to 29.5 million working age adults. By comparison, the number of persons receiving disability benefits under DI and SSI during the

---

[16] See McNeil, 1993.

Table 6.—Selected estimates of the number and percentage of the working age population[1] with disabilities[2] and the number of adults receiving Social Security or Supplemental Security Income benefits based on their own disability, late 1991

Disability Definitions	Persons (in thousands)	Percent of total
Total persons	165,040	100.0
With a disability[3]	29,482	17.9
With a severe disability[4]	13,171	8.0
With a work disability and prevented from working[5]	7,588	4.6
With one or more severe functional limitations	6,596	4.0
Has difficulty performing one or more activities of daily living	3,442	2.1
Total receiving DI or SSI disability benefits	5,702	3.5
DI beneficiaries[6]	3,878	2.3
SSI recipients[7]	1,824	1.1

[1] Aged 15-64 except as noted.

[2] As estimated from the 1990 and 1991 panels of the Survey of Income and Program Participation. Interviews took place from October 1991 through January 1992.

[3] Persons identified as: (1) having a work disability or housework disability; (2) having difficulty in performing one or more of six functional activities, six activities of daily living (ADLs), five instrumental activities of daily living (IADLs); (3) using a wheelchair, crutches, a cane, or walker; (4) having one of five classes of impairments (a learning disability such as dyslexia, mental retardation, other developmental disability such as autism or cerebral palsy, Alzheimer's disease/senility/dementia, or any other mental or emotional condition); or (5) a nonaged beneficiary of the Medicare or SSI program.

[4] A person identified as: (1) prevented from working at a job or business or from doing work around the house; (2) unable to perform 1 of 17 functional, ADL, or IADL activities; (3) having one of three classes of impairments (mental retardation, a developmental disability, or Alzheimer's disease/senility/dementia); or (4) using a wheel chair or a long-term user of crutches, a cane, or walker.

[5] Aged 16-64.

[6] Number of Social Security beneficiaries, aged 18-64, receiving benefits as of December 1991 based on their own disabilities, for example, disabled workers, disabled widows and widowers, and disabled adult children.

[7] Number of adults aged 18-64 receiving a federally administered payment but not also receiving DI benefits as of December, 1991.

Source: John McNeil, "Americans with Disabilities: 1991-92," U. S. Bureau of the *Census, Current Population Reports, Series P-70-33,* U.S. Government Printing Office, Washington, DC, 1993. Beneficiary and recipient estimates based on SSA administrative data.

A:\DISCNTS.XLS   8/17/94 rev.: 2/23/96

period was about 5.7 million. Hence, depending on the definition of disability selected, the number of working age disabled can be up to five times the number of beneficiaries.

One Approach: Estimating a Model of Disability Determination with Matched Data
Although how survey responses relate to disability determinations is not self evident, it is a relationship that can be modeled statistically provided the necessary data are available. In the end, efforts to use household survey data to study disability programs depend on a central methodological question: How, if at all, can survey responses relating to health, demographic factors, activity limitations, and work be used to identify those who would be considered disabled under SSA criteria?

Recent research at the Social Security Administration has addressed this issue using SIPP data exact-matched to SSA records on disability determinations.[17] This study models the outcomes of SSA disability determinations for adult applicants, using as explanatory variables survey responses on health, functional limitations, demographic traits, and work experience. These variables are shown to be systematically related to the determinations of SSA adjudicators and the relationships estimated are plausible to those familiar with the disability determination process. This effort will support estimation of the number of persons in the general population eligible for the disability programs. Beyond that, it will allow development of a comprehensive model involving both government decisions (that is, eligibility determination) and individual behavior (that is, applications decisions), to explain program growth. This approach will allow us to distinguish the effects of categorical criteria, financial criteria, and the application incentives faced by households.

Table 7 deals with the data and modeling requirements for program analysis of the DI program. DI is doubly challenging in that it poses obstacles in terms of both the categorical

---

[17] See Lahiri, Vaughan, and Wixon (1995). All matching activities for this study were carried out as part of a joint SSA-Bureau of the Census statistical project under the aegis of the agencies' Memorandums on the Exchange of Statistical Information and Service. All work involving the development and analysis of the matched data set at SSA was carried out by SSA employees (or on-site contractors) acting as special sworn agents of the Bureau of the Census.

Table 7

Disability Insurance Benefits: Matched Survey Data, Determination Model, and Benefit Simulation

<u>Populations</u>

<u>Attributes</u>	Program Participants	Nonparticipants	
		Eligible	Ineligible
Program Eligibility Variables			
Categorical Eligibility (health, activity limitations, . . .) *Financial Eligibility* (earnings history, quarters covered, family comp., marital history . . .)	**Program Data**	(Eligibility Simulated)	(Eligibility Simulated)
Evaluative Variables			
Well Being (income, assets . . .) *Contributions* (earnings history, taxes) *Demographic Traits* (age, race, ethnicity, education, health . . .) *Household Events, Composition* (marital status, loss of spouse, job loss, loss of health ins. . . .)	**Matched Survey Data (with Disability Determination Model, Benefit Simulation)**		

and financial criteria for eligibility, necessitating matching of both disability determination data and earnings history data. The problem of categorical eligibility, discussed in this section, also applies to the disability portion of the SSI program.

Neither program data nor household surveys--when used in stand alone mode--support a basic course of program analysis for the disability programs. Program data provide a wealth of detail on the impairment, SSA's determination and the criteria for it, earnings, the period working in a job covered by social security and so on. But, once again, such data do not allow analysis of the program's distributional effects in terms of comprehensive measures of well-being or in terms of a broad set of demographic traits. Nor do program data include information on, for example, recent household events, information needed for behavioral analysis. And, while household surveys have a full array of attributes, they lack necessary program detail. Without data matching and eligibility modeling, household surveys allow only analysis exploiting the most basic program category--whether the survey respondent receives benefits or not.

The administrative data on disability determinations serve two functions in this modeling effort. First, they permit identification of survey sample members who applied for benefits during the period proximate to the survey, including those denied as well as those allowed. Second, they enable separate analyses of applicants allowed or denied under different decision criteria. For example, some applicants are allowed because their impairments met SSA's medical criteria (called the medical listings), while for others the decision takes into account not only medical criteria, but age, education, work experience, and an assessment of the applicant's capacity for work. Combined, these two alternative criteria provide the basis for the allowance decisions of the eight million persons currently receiving benefits. The administrative data permit distributional and behavioral analyses of each group separately, whereas survey information would permit only analyses of "composite" effects of the two distinct criteria (that is, analysis of all disabled beneficiaries together). This illustrates how matching allows analysts to focus on specific normative criteria underlying eligibility and benefit decisions. Moreover, the same analytical opportunities would not be available by designing better surveys, since survey respondents would not know under which of the two

criteria they receive benefits.

6. Conclusions

There is a data requirement central to the analysis of all public programs considered here: What is needed to support microanalytic study is a data set which combines evaluative variables (such as measures of well being, demographic characteristics, household composition, and household events) with detailed programmatic information (such as program eligibility, benefit size, and how an individual's eligibility or benefit is affected by a specific program provision). Juxtaposing such attributes for individuals represented in the data set is basic to analysis of how the program is targeted and the incentives it poses.

Juxtaposing these two types of attributes, however, is problematic. Neither survey data nor program data, taken alone, offer both. Moreover, these limitations are not accidental--they reflect basic features of administrative data collection and survey data collection. While what I have called evaluative variables are frequently collected in surveys, few are collected in the course of program administration, since, by statute and regulation, most such variables have no role in determining eligibility or benefits. In addition, administrative data have detailed program information, but only for participants. Conversely, while surveys collect a more comprehensive set of evaluative variables, in terms of program information, they often ask about benefit receipt and size only. More detailed program information (such as the eligibility of nonapplicants or whether affected by specific benefit criteria) are likely uncollectable using standard household surveys.

Fortunately, a reliable means of replicating government decisions on eligibility and benefits has been in use for over two decades. Eligibility models, sometimes referred to as microsimulation models, can be used to simulate eligibility and benefit decisions, often using surveys which have collected the necessary elements. Such models estimate both categorical eligibility (such as advanced age or single parent status) and financial eligibility. Because such models can be used to add program information to the attributes collected in surveys,

they represent a linchpin for program analysts. That is, such surveys can only be fully exploited for distributional and behavioral analysis if program eligibility can be simulated. This implies, in turn, that basic program analysis using household surveys depends on whether the survey, matched or unmatched, meets the data requirements for eligibility modeling.

I have illustrated that for the SSI (aged) program--considered as an example of means-tested programs generally--eligibility and benefits can be simulated using information collected in household surveys having comprehensive measures of financial resources. And, as a consequence, household surveys, without administrative data matching, can support modeling of program participation decisions and simulation of many alternative benefit structures, for example. This success story follows from a coincidence between principal features of means-tested programs and household surveys. Means-tested programs use family financial status in the current period as a basic criterion, while for cross sectional surveys, such as the SIPP, the financial well being of the family just prior to the survey represents a primary focus. As a result, such surveys include the elements of eligibility for most means-tested programs.

But household surveys do not confer equivalent analytical opportunities for all public programs. Contributory programs have particularly vexing data requirements. In the case of social security, the obstacle is that surveys do not include lifetime earnings, which are needed for benefit simulation. This precludes much retirement modeling. It also prevents distributional analysis, such as evaluating program provisions having adequacy objectives in the light of the family income of those affected. Matching of earnings data from administrative sources allows simulation of the social security benefit, enabling much basic analysis of program targeting and applications behavior. By comparison, analysts of means-tested programs can undertake an analogous course of study using unmatched household surveys.

Finally, there is a separate problem unique to SSA's disability programs. On the basis of information collected in surveys, how can analysts estimate who would be found disabled under the criteria used by SSA? The inability to estimate those eligible in terms of categorical (nonfinancial) criteria has limited the use of household data in analyzing program

growth. One solution is presented here. SSA analysts have matched administrative information on disability determinations to household survey data; as a result, the outcome of SSA determinations can be modeled using survey responses on health, activity limitations, demographic traits, and work behavior. This approach will allow analysis of both government decisions (that is, eligibility determination) and individual behavior (that is, applications decisions) in a framework that distinguishes the effects of categorical criteria, financial criteria, and the incentives faced by households.

References

Advisory Council on Social Security (1979).
   **Social Security Financing and Benefits: Report of the 1979 Advisory Council on
   Social Security.**

Kilss, Beth and Frederick J. Scheuren (1978). *The 1973 CPS-IRS-SSA Exact Match Study,*
   Social Security Bulletin, 41:10:14-22.

Lahiri, Kajal, Denton R. Vaughan, and Bernard Wixon (1995). *Modeling SSA's Sequential
   Disability Determination Process Using Matched SIPP Data,* **Social Security Bulletin,**
   58:4:3-42.

McClung, Nelson (1986).
   The Art of Transfer Policy Analysis, in G. H. Orcutt, J. Merz, and H. Quinke, editors,
   Microanalytic Simulation Models to Support Social and Financial Policy. North-
   Holland.

McNeil, John M. 1993. *Americans with Disabilities: 1991-92, Data from the Survey of
   Income and Program Participation.* U.S. Bureau of the Census, Current Population
   Reports, Series P-70, No. 33, Washington, DC: U.S. Government Printing Office,
   December.

Projector, Dorothy S. and Ellen G. Murray (1978). *Eligibility for Welfare and Participation
   Rates, 1970.* **Studies in Income Distribution,** number 7.

Social Security Administration. 1995. *Annual Statistical Supplement* to the *Social Security
   Bulletin.* U.S. Government Printing Office, Washington, DC.

Vaughan, Denton R. and Bernard Wixon. 1991. "Two Papers on a New SIPP-Based
   Microsimulation Model of SSI and OASDI." ORS Working Paper Series, Number
   54, Division of Economic Research, Office of Research and Statistics, Social Security
   Administration, Washington, DC.

Wixon, Bernard, Benjamin Bridges, Jr. and David Pattison (1987).
   Policy Analysis Through Microsimulation: The STATS Model, **Social Security
   Bulletin,** 50:12:4-12.

December 11, 1996
wixon-c:\usrfiles\wp\model\match5.ppr

DISCUSSANT'S REMARKS
Using Administrative Records for Statistical Purposes

Before getting into the substance of the two papers, let me say a little about my own work, which will indicate how sympathetic I am to the positions of the two authors. I am an economist at the Social Security Administration whose research concerns the interactions between Social Security's programs and the labor market. That is, how do American work and retirement patterns affect Social Security revenues and expenditures and, in turn, how do our programs influence work patterns and trends. Some of my research uses the matched SIPP files to which Bernie Wixon has referred. In addition, for several years I worked on a project with a handful of people at the Census Bureau and SSA to match our administrative data to a data set we haven't mentioned here: the National Longitudinal Survey of Women. Just this year we have finally created this matched file, which is the primary data resource being used in research we are conducting on the Social Security program's treatment of women.

Let me say a little more about this project to supplement the discussion in the two papers we've heard. In recent years a number of policy questions have arisen regarding women and the Social Security program. Just a few of these:

- How does Social Security treat women who make different choices regarding marriage and work? For example, a woman who works her entire life and pays substantial Social Security taxes could end up receiving smaller Social Security benefits than a married, full-time homemaker who pays little or no Social Security taxes and receives benefits on the basis of the earnings record of a high-earning husband. Over time have spousal and survivor benefits paid to women been increasingly paid to higher income families where wives can afford not to work outside the home?

- Although the poverty rate for persons aged 65 or older has fallen substantially in the past 3 decades, and is now lower than the rate for the population aged 18-64, the rate for older women -- particularly those who live alone -- remains high. Are there appropriate Social Security policy changes that might remedy

347

this?

- Do labor market trends in hours and earnings for women suggest that their lifetime earnings patterns will eventually closely resemble those of men? How will the type and size of women's benefits differ in the future? And, what are the consequences for the long term finances of the Social Security program?

When we began to think about a research program to address these types of questions, an attractive data source seemed to be the National Longitudinal Survey of Women. Back in 1967 the Bureau of Labor Statistics began interviewing two samples of women: the NLS Mature Women who were then aged 30-44, and the NLS Young Women who were then aged 13-21. These samples have been regularly interviewed ever since about their work experiences, education and training, family backgrounds, child rearing activities, health status, incomes, living expenses, saving and investment, retirement planning, and a host of other information relevant for explaining their eventual economic status in old age. Of course, we would like to see precisely how these women do and will interact with the Social Security program. The Mature Women are now in their 60s and 70s and for the most part have reached their retirement years. The Young Women are in their 40s and we can begin to see how their generation will bring quite different earnings and marital histories into their retirement years.

Just this year we have linked the Social Security Administration's record data on earnings and benefits to the NLS survey data for these two groups of women (and their husbands) which enables us to explore a wide range of policy-relevant research questions. For example, what factors explain the incidence of poverty for older women? Are younger women preparing for retirement any better than the previous generation? How are the type and amount of Social Security benefits likely to change as successive cohorts of women retire? Are the determinants of the timing of retirement for women the same as for men, who have been much more extensively studied? Most of these types of questions cannot be pursued

without survey data linked with the Agency's administrative data. On their own, surveys -- even those as comprehensive as the NLS Women -- do not have the accurate Social Security covered earnings information needed to compute the basis and amount of a woman's retirement benefit. And, the Agency's administrative data lack nearly all the information that social scientists require to explain behavior. That is, what explains a woman's lifetime earnings pattern, why did she apply for benefits when she did -- or when is she likely to -- and how important is her Social Security benefit to her economic wellbeing? A matched survey file of this type offers enormous potential for policy research.

Now, on to the two papers.

Specific remarks on Ron Prevost's paper:

Ron Prevost has laid out many of the important questions that will have to be addressed if a variety of administrative record files are to be used as key inputs into the decennial census planning and information gathering. Our experience at SSA with trying to use our own administrative files suggests that Ron's research agenda will take a small army to implement. Administrative data are often ambiguous, inaccurate, awkwardly managed, and poorly documented -- if documented at all. Keep in mind that these shortcomings do not necessarily impede an agency from carrying out its basic administrative functions effectively. Rather, they simply demonstrate that recordkeeping procedures that might work for program administrators can impose thorny problems for the researcher wanting to use these data. It will no doubt be difficult to access, interpret, and merge information from many different administrative sources.

These concerns based on the experience of my research colleagues at Social Security lead me to pose this question. Keep in mind that I'm an economist, not a survey statistician. My understanding is that survey methodologies are now so sophisticated that sampling

techniques allow one to get a more accurate estimate of population characteristics than efforts to count the population in an exhaustive manner, as in the decennial census. If the decennial census has problems due to escalating costs and an uncooperative public, is greater reliance on administrative data from many sources the most cost-effective solution? Why this option as opposed to a greater reliance on sample surveys? I would be curious to hear the answer.

Specific remarks on Bernie Wixon's paper:

I am well acquainted with the disability research project that serves as a basis for many of the ideas in my colleague Bernie Wixon's paper. This is an excellent example of how administrative data linked to survey information can open up a set of questions for investigation that were otherwise impossible to address.

In many instances an Agency's administrative data will enable a researcher to create a statistical description of specific elements of an Agency's operations that cannot be accomplished via any alternative means. But, matched data can also be used to improve the scope and quality of behavioral research, as Bernie's paper has emphasized. Think of what a researcher can learn from an Agency's administrative files. Often we can recover key data elements from administrative forms that were processed. We can review applications, administrative decisions, outcomes, payments, and the like. This is all valuable stuff and can be illuminating in its own right. But, if we are interested in questions such as: Why did this person submit an application, and why then? Why did other individuals with ostensibly similar circumstances not apply? Will particular changes in our program rules redistribute benefits to financially needy families? Would that new policy cause people to behave differently? Administrative files rarely contain the necessary information. In contrast, what can a researcher learn from household surveys such as the CPS, SIPP, and PSID files, but without access to an agency's administrative files? Often there's plenty of background information on

the demographic and socioeconomic characteristics of individuals and their families. But, from the perspective of policy research and analysis, there is usually insufficient detail about how the respondents interacted with a specific government program. When survey and administrative data are matched, the result is an information base that can support considerably improved research projects that assess the effectiveness of current and proposed government programs.

Although there are clearly grounds for optimism on this score, even with the creation of high quality matched survey data sets, the research problems are formidable. The centerpiece of behavioral social science research is usually the statistical estimation of some model intended to represent and explain some behavioral phenomenon. So, one might hypothesize a simple model with general functional form:

$$y = f(X, \beta, \varepsilon)$$

where y is the behavior to be explained or predicted, X is a set of explanatory factors, β is a vector of coefficients or parameters to be estimated, and ε is an error term that specifies the relationship's stochastic properties. The objective is to estimate the true values of β by applying appropriate statistical methods to our data.

The list of the main things that can go wrong in this process is collectively known as *specification error*. Specification error can be of four basic types:

1. Incorrect functional form. The researcher usually has to make decisions and choices about f(.) and it is very easy to get things wrong. Social science theories -- even economic ones -- have little specific to offer on this count. In the large number of cases where we really don't have much of a theory at all, it is very easy to wander far astray at this point.

2. Incorrect set of explanatory variables. These are of two types of mistakes: including variables that don't belong and omitting ones that do. It is easy to include

factors that seem plausible but simply don't belong, thereby imposing inefficiency in the estimation. Even worse is the case of leaving out explanatory variables that belong in the model, thereby suffering from inefficient estimation as well as admitting bias into the estimation of the model's parameters. The main reasons for omitting variables that belong in our model are flawed theories and data sets that simply have not collected the relevant information -- in some cases because the phenomena are unobserved or unmeasurable (*e.g.*, motivation, tolerance for pain).

3. Errors in variables. The explanatory variables are measured with error. For example, we might think that income is involved in our story, but the survey has measured income inaccurately.

4. Incorrect specification of the model's error term. Empirical researchers can be remarkable cavalier about this aspect of model development, leaving the stochastic specification as a maintained hypothesis. The ease of using today's canned statistical packages makes this error extremely easy to commit.

All of these forms of specification error potentially invalidate the model results and

their interpretation. So, what contribution do matched surveys make to this process? They

help potentially with the second and third types of problems. Sometimes the addition of

administrative data allows a researcher to include information in the model that is available

from no other source. An example might be the absence of a worker's earnings history in a

survey such as SIPP, that can be obtained from SSA's administrative records. Or, sometimes

we have imprecise measurement of certain explanatory factors that can be improved by

consulting an Agency's records. For example, during what time interval in the past did an

individual receive a particular type of government transfer and in what amount? This is the

type of information that is frequently inaccurately reported in surveys.

In any event, the availability of matched data offers the promise of improving the

quality of policy research on behavioral issues. In light of the daunting problems that credible

research of this type must surmount, it falls considerably short of a panacea.