

Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project

John M. Abowd, Martha Stinson and Gary Benedetto

November 5, 2006

This report was produced by the Longitudinal Employer-Household Dynamics Program at the U.S. Census Bureau, Jeremy S. Wu, Assistant Division Chief, Data Integration Division. The report is required by the Jointly Financed Cooperative Agreement between the Census Bureau and the Social Security Administration for fiscal year 2006 (SSA agreement number BC-05-05, as amended; Census Bureau agency reference number 0084-2005-043-002-001, project account 7675084). John Abowd participated in the project in his capacity as Distinguished Senior Research Fellow at the Census Bureau (on IPA from Cornell University). Martha Stinson and Gary Benedetto are economists on the LEHD staff. In addition to the authors named above, Lisa Dragoset (Census-LEHD), Sam Hawala (Census-SRD), Karen Masken (IRS), Bryan Ricchetti (Census-LEHD), Lars Vilhuber (Census-LEHD), and Simon Woodcock (Simon Fraser University) all contributed to the research. Josep Domingo-Ferrer (University of Rovira and Virgili), Jerome Reiter (Duke University), Vicenc Torra (Artificial Intelligence Lab, University of Barcelona), and Simon Woodcock, operating with the support of the Census Bureau through a subcontract to the main Research and Development contract between the Census Bureau and Abt Associates, Inc. (Census Bureau contract number 50YABC-2-66036, task order number TO002) to Cornell University (OSP reference number 47632), provided substantial consulting on the creation of the SIPP/SSA/IRS-PUF. The National Science Foundation through Grants ITR-0427889 and SES-0339919 to Cornell University with subcontracts to the Census Bureau (Census Bureau agreement number 0063-2005-003-000-000, project account 9098000) also provided substantial support for this project.

Contents

1	Executive Summary	1
1.1	Purpose and brief history	1
1.2	Structure of the inputs to the SIPP/SSA/IRS public use file	1
1.3	Completion of the missing data and synthesis of the confidentiality-protected data	2
1.4	Development of the weights	3
1.5	Analytical validity testing	4
1.5.1	All univariate distributions	5
1.5.2	Summary statistics for all workers and for OASDI beneficiaries	5
1.5.3	Summary statistics by education and foreign born	5
1.5.4	Selected regression model results	6
1.6	Disclosure avoidance assessment	6
1.7	Using the SIPP/SSA/IRS-PUF	7
1.8	Next steps	7
2	Project Background	9
2.1	Purpose and brief history	9
2.2	Overview project description	9
3	Creation of the Gold Standard File	11
3.1	SIPP data	11
3.2	IRS/SSA earnings data	13
3.3	SSA data	13
3.4	Weight creation and use	14
3.5	Gold Standard data dictionary	14
4	Data Completion and Synthesis	15
4.1	General methodology	15
4.2	Bayesian Bootstrap	17
4.2.1	Generic BB algorithm	17
4.2.2	Bayesian bootstrap application	18
4.3	Sequential Regression Multivariate Imputation	19
4.3.1	Definitions and general algorithm	19
4.4	Summary of synthetic data production	20
4.5	Modeling details	22
4.5.1	Types of variables	22
4.5.2	Parent-child relationships and constrained variables	24
4.5.3	Levels of the parent/child tree	25
4.5.4	Grouping and conditioning variables	25
4.5.5	Specific variable details	26
5	Weight Creation and Synthesis	32
5.1	Introduction and background	32
5.2	Summary of the weight creation process	33
5.2.1	Part A: Creation of poverty stratification variable for Census 2000 records	33
5.2.2	Part B: Creation of stage-2 clusters for Census 2000 records	33
5.2.3	Part C: Creation of poverty stratification variable for SIPP records	33
5.2.4	Part D: Creation of stage-2 clusters for SIPP records	33
5.2.5	Part E: Matching SIPP individuals to the Census 2000 records	33
5.2.6	Part F: Creation of a preliminary weight	34
5.2.7	Part G: Creation of final weight	34

5.3	Part A: Creation of poverty stratification variable for Census 2000 records	34
5.3.1	Data sources for short-form respondents	34
5.3.2	Data sources for long-form respondents	34
5.3.3	Data source for MSA variable	35
5.4	Poverty stratum assignment	35
5.5	Part B: Creation of stage-2 clusters for Census 2000 records	36
5.5.1	Creation of PSUs	37
5.5.2	Creation of stage-1 clusters	37
5.5.3	Creation of stage-2 clusters	37
5.5.4	Dropping Census 2000 records that were out-of-scope for SIPP samples	37
5.5.5	Census 2000 stage-2 cluster tabulations	38
5.6	Part C: Creation of poverty stratification variable for SIPP records	38
5.7	Part D: Creation of stage-2 clusters for SIPP records	38
5.8	Part E: Matching SIPP individuals to Census 2000 records	39
5.8.1	Un-duplication of SIPP-Census 2000 matches	39
5.8.2	Matching SIPP to Decennial through probabilistic record linking	39
5.9	Part F: Creation of preliminary weight	41
5.10	Part G: Creation of the final weight	41
5.11	Geography issues	41
5.11.1	Different geography concepts on HCEF and other Census 2000 files	41
5.11.2	Changing geography boundaries between the 1990s and 2000	42
5.12	Birth date issue	42
5.13	Overall evaluation of Gold Standard weight	42
5.14	Synthesizing the weight	42
5.15	Evaluation of the synthesized weight	43
6	Analytical Validity	45
6.1	Complete data estimation	45
6.2	Inference frameworks using multiple imputation	46
6.2.1	Missing data only	46
6.2.2	Missing and partially synthetic data	47
6.3	Application to the SIPP/SSA/IRS-PUF	48
6.4	Results	49
6.4.1	General interpretation	49
6.4.2	Summary statistics for OASDI beneficiaries	49
6.4.3	Summary statistics for all workers	50
6.4.4	Summary statistics by education categories	50
6.4.5	Summary statistics by foreign born	51
6.4.6	Summary statistics for marital histories	51
6.4.7	Age at time of retirement	52
6.4.8	Selected regression results	52
6.4.9	Univariate distributions of continuous variables	54
6.4.10	Counts and percentages of categorical variables	55
7	Assessing Disclosure Risk	56
7.1	Overview	56
7.2	Matching based on probabilistic record linking	56
7.3	Distance matching	58
8	Using Synthetic Data	61
9	Conclusion	63

Bibliography

64

A Appendix

65

1 Executive Summary

1.1 Purpose and brief history

The creation of public use data that combine variables from the Census Bureau's Survey of Income and Program Participation (SIPP), the Internal Revenue Service's (IRS) individual lifetime earnings data, and the Social Security Administration's (SSA) individual benefit data began as part of ongoing collaborative research at the Census Bureau and SSA. The current project had its genesis with the formation of a joint committee containing representatives from the Census Bureau, SSA, IRS, and the Congressional Budget Office (CBO) that designed a prospective public use file. Aimed at a user community that was primarily interested in national retirement and disability programs, the selection of variables for the proposed SIPP/SSA/IRS-PUF focused on the critical demographic data to be supplied from the SIPP, earnings histories from the IRS data maintained at SSA, and benefit data from SSA's master beneficiary records.

After attempting to determine the feasibility of adding a limited number of variables from the SIPP directly to the linked earnings and benefit data, it was decided that the set of variables that could be added without compromising the confidentiality protection of the existing SIPP public use files was so limited that alternative methods had to be used to create a useful new public use file. The committee agreed to allow the Census Bureau to experiment with the confidentiality protection system known generically as "synthetic data." The actual technique adopted is called partially synthetic data with multiple imputation of missing items. As the term is used in this report, "partially synthetic data" means the release of person-level records containing some variables from the actual responses and other variables where the actual responses have been replaced by values sampled from the posterior predictive distribution for that record, conditional on all of the confidential data.

From 2003 until the present, four preliminary versions of the SIPP/SSA/IRS-PUF have been produced. This final report accompanies the delivery of version 4.0 to SSA as part of the fiscal year 2006 Jointly Financed Cooperative Agreement between the Census Bureau and SSA.

1.2 Structure of the inputs to the SIPP/SSA/IRS public use file

The SIPP/SSA/IRS-PUF contains data from the records of individuals who responded to the SIPP panels conducted in 1990-1993 and 1996. A standardized extract of approximately 125 variables from all waves of each of these panels was created. We included the following demographic variables: gender, marital status, race (black), five categories of education, Hispanic ethnicity, birth date, death date, disability status, number of children, marital history, foreign born, decade arrived in United States if foreign born, and a spouse identifier that links to the marriage partner if the respondent is married and the spouse was also surveyed. We took the values for these variables at a point in time. For the time-invariant variables—gender, race, and Hispanic ethnicity, values were taken from the point in the SIPP when they were first reported, generally wave 1. Values for the other demographic variables were generally chosen from month 8 of the respective SIPP panel (*i.e.*, the last reference month of the second interview). We chose this point because marital, immigration, and disability histories were collected in the wave 2 topical modules and we wanted to take all the variables from the closest possible interview dates. For education, we searched over all reported education values in each wave of the SIPP and chose the highest level of education ever reported. Thus gender, marital status, race, education, Hispanic ethnicity, and spouse identifier are never missing in the standardized extract because these variables are all reported at least once, and we chose to take the self-reported values whenever they were available. Disability status, number of children, marital history, foreign born, and decade arrived in United States if foreign born are sometimes missing because individuals did not answer the relevant topical modules or because we chose not to search over every available month of SIPP data. All item missing data, with the exception of structurally missing items, were flagged for imputation.

This standardized extract was linked using the respondent's validated Social Security Number (SSN) to the following data provided by SSA:

- From SSA's Summary Earnings Record (SER), a longitudinal history of all FICA-covered wage and salary income earned since 1937, we linked the annual summary and the quarters-worked summary. These are the only earnings data available from the SSA and IRS files prior to 1978. This array is capped at the FICA taxable maximum;¹

¹These data, as well as the Detailed Earnings Record data cited in the next bullet, are also confidential under the protocol defined in Title 26 of the U.S. Code. Prior permission from the IRS disclosure officer is required before they can be used in a project in combination with Title 13

- From SSA’s Detailed Earnings Record, a longitudinal history of wage and salary items from the employer-filed W-2 form by employer, we linked annual total wage and salary income and deferred earnings from all FICA-covered jobs. We also linked an analogous set of variables for non-FICA-covered jobs;
- From SSA’s Master Beneficiary Record (MBR), a longitudinal history of type and amount of all benefits paid to an individual, we linked the entire history and created variables for type of benefit initially received, type of benefit received in April 2000, and the monthly benefit amount associated with those two benefit receipt dates.
- From the Census Bureau version of the master Social Security Number data base, known as Numident when sorted in SSN order, we linked the administrative birth and death dates.

Next, we added variables that were not destined for the public use file but would provide additional information useful in the process of completing the missing data, synthesizing the variables to be protected, creating a weight for the merged SIPP panels, and assessing the quality and disclosure risk of the final product. The documented, standardized extract from the SIPP 1990-1993 and 1996 panels, the linked SSA and IRS data, the supplemental variables added to facilitate processing and review, and the customized weight collectively define what we call the “Gold Standard” file. The codebook and technical description of the Gold Standard Version 4.0 accompanies this report. This codebook also documents the variables found in the completed Gold Standard files and the synthetic data files.

1.3 Completion of the missing data and synthesis of the confidentiality-protected data

Although the existing SIPP public use files have had all item non-response allocated using the methods developed for this purpose as part of the regular SIPP data processing, the Gold Standard version of the consolidated 1990-1993 and 1996 panels has item missing data for two basic reasons. First, SIPP respondents in the Gold Standard file for whom the Census Bureau does not have validated SSNs were missing all data items whose linkage depends upon the SSN; that is, all earnings, benefit, and administrative birth and death data. Second, because one of the critical components of the confidentiality protection is to prevent identifying the source record of the synthetic data in the existing SIPP public use files, all information regarding the dating of variables whose source was a SIPP response, and not administrative data, has to be made consistent across individuals regardless of the panel and wave from which the response was taken. This requirement resulted in the creation of ten-element arrays that contained all dated SIPP items, like family total income, with values inserted for each year from 1990 to 1999. No SIPP respondent household ever provided all ten of these items. Those array elements that were available for a particular respondent, which depend upon which panel the respondent answered, were populated by the actual value (from the public use version of the variable). All other elements in the array were item missing data. All missing data items that resulted from either missing validated SSN or missing items in an array were multiply imputed using the techniques described in the report. The imputation models were based on Bayesian bootstrap and Sequential Regression Multivariate Imputation methods for estimating and sampling from multivariate posterior predictive distributions.

There is a third source of missing data in the Gold Standard file. Some data items are structurally missing because it is not logically possible for the item to have a value; for example, no data are available concerning the second marriage of individuals who never married or married only once. Structurally missing data remain in the Gold Standard file and in the synthetic data implicates that constitute the SIPP/SSA/IRS-PUF.

The public use file contains several variables that were never missing and are not synthesized. These variables are: gender, marital status, spouse’s gender, initial type of Social Security benefits, type of Social Security benefits in 2000, and the same benefit type variables for the spouse. All other variables in the public use file were synthesized.

In order to preserve exact logical relations among the variables, the first step of the missing data imputation process, and the first step of the data synthesizing process, is to implement a binary tree of parent-child relations among all the variables. This tree guides the execution of first the missing data imputation and then the synthetic data phase. We created the binary tree to organize the data processing by summarizing all of the assumptions and logical restrictions that must be preserved in the final data product.

The top level of this binary tree contains all variables that exhibit no logical dependencies on any other variables in the file, for example birth date. The tree has nine levels. At each level below the top, variables depend upon their

confidential data. Permission to conduct the present research is monitored by the Census Bureau under Administrative Records Tracking System project 458, which contains a copy of the IRS approval.

parents, and are only processed when appropriate. In the intermediate levels of the tree, a variable can be both a parent and a child, for example, whether or not there is a second marriage is a child of the same variable for the first marriage and a parent of the variable for the third marriage. The terminal level and all leaves of the binary tree contain only child variables.

For each iteration of the missing data imputation phase and again during the synthesis phase, we estimate a joint posterior predictive distribution for all of the required variables according to the following protocol. At each node of the parent/child tree, a statistical model is estimated for each of the variables at the same level. The statistical model is a Bayesian bootstrap, logistic regression, or linear regression (possibly with transformed inputs). All statistical models are estimated separately for detailed groups of individuals based on the values of categorical variables that include both demographic and economic controls. Logistic and linear regressions also include additional linear controls that are selected from a long list of potential right-hand-side control variables on the basis of the Bayes Information Criterion. Once the analyst specifies the grouping variables and their associated control variables, the estimation of a proper posterior predictive distribution from which to impute or synthesize, as appropriate, is fully automated. On the basis of the estimated models, and taking proper account of parameter uncertainty, each variable is imputed (missing data phase) or synthesized (synthetic data phase) conditional on all values of all other variables for that individual. The missing data phase included nine iterations of estimation. The synthetic data phase occurred on the tenth iteration. Four missing data implicates were created. These constitute the completed data files that are the inputs to the synthesis phase. Four synthetic implicates were created for each missing data implicate. Thus, there are a total of sixteen synthetic implicates in the SIPP/SSA/IRS-PUF Version 4.0.

A complete diary of the assumptions used to synthesize every variable in the PUF: parent/child relations, synthesizer method, statistical model, grouping variables, control variables, allowable values, logical limitations, synthesizer restrictions, and usage notes is included as an Excel workbook accompanying this report.

The software to implement the missing data imputation and confidentiality synthesis is written in SAS as a massively parallel application. Running on two 64-processor large memory computers at the Census Bureau the estimation phase for completing all 616 variables can be accomplished in about two months. Given completed data, a full run of the synthesizer (16 implicates) takes about three weeks.

1.4 Development of the weights

The final Gold Standard file contains data drawn from the survey responses and administrative records of individuals who responded to the Survey of Income and Program Participation in the 1990-1993 and 1996 panels. The design of the 1990-1993 panels envisioned combining data from waves of different SIPP panels that corresponded to the same calendar dates. Consequently, there are explicit instructions for recalibrating the SIPP weights when using individuals or households from the same year who were surveyed in different panels. The recalibrated weights account for the design and ex-post differences across the panels. The data collected as part of the 1996 panel do not overlap the time periods covered by the 1990-1993 panels. Hence, no official formulae exist for recalibrating the weights when combining data from the 1996 panel with data from the earlier panels.

The linkage of longitudinal lifetime earnings data from SSA's Summary and Detailed Earnings Records to individuals from these five SIPP panels implies that records that correspond to the same calendar year will come from all of the panels. Analyses that use these longitudinal earnings data cannot use any combination of the official SIPP weights to produce an estimate that has a fully specified reference population. This conundrum has faced analysts who used linked SIPP/SSA/IRS data, such as internal researchers at SSA and the Census Bureau, for years. In order to allow users of the SIPP/SSA/IRS-PUF Version 4.0 to conduct analyses with a known reference population, we created an ex post weight for the PUF. This weight can be used to make estimates representative of individuals age 18 or older in the civilian non-institutionalized U.S. population as of April 1, 2000, the reference date for Census 2000.

Our method for creating an ex-post weight for the merged SIPP panels involved seven steps. First, we reproduced the major component of the 1996 sampling frame (the unit frame) in the Census 2000 micro-data, updating the SIPP reference population to April 1, 2000. Next, we divided the Decennial individual records (long and short form) into strata according to the 1996 SIPP sampling plan. Then we standardized all five SIPP panels with respect to geographic definitions and strata used in the 1996 sampling plan. Each individual observation in the merged panels was then placed in a stratum according to the 1996 SIPP sampling plan. The fifth step was to link each SIPP person to a person in the Census 2000 reference population. This match was accomplished using probabilistic record linking. Most observations

could be linked on the basis of the PIK² that had been assigned to the SIPP or Decennial individual. For those SIPP individuals who did not link by PIK, a cruder probabilistic record linkage based on characteristics used to define the sampling strata was used. Having accomplished this linkage for all in-scope individuals in the 1990-1993 and 1996 SIPPs, we created a preliminary weight as the ratio of in-scope individuals in Census 2000 to in-scope individuals in the merged SIPPs within each final (stage-2) SIPP sampling cluster. The final weight was created by raking the preliminary weights to agree with official U.S. population control totals for the sex/age/race/ethnicity demographic breakdown of the reference population, as supplied by the Census Bureau's Population Estimates Division. This final raking was controlled to exactly the same population categories as the official 1996 SIPP weights.

The final weight was tested for analytical validity by creating weighted tables summarizing earnings and benefit measures from the administration of the Old Age, Survivor, Disability Insurance (OASDI) program. The estimates from the PUF were compared to SSA's published statistical summaries for the year 2000. When the final weight is applied to the completed Gold Standard data, the results reproduce most aspects of published SSA data derived from the universe of OASDI recipients.

Because copying the final weight to each implicate of the synthetic data would have provided an additional un-synthesized variable with 55,552 distinct values, the disclosure risk associated with the weight variable had to be addressed. We created a synthetic weight using a posterior predictive distribution based on the Multinomial/Dirichlet natural conjugate likelihood and prior. The likelihood component was created by modeling the 55,552 distinct cells created by all feasible combinations of the six variables used to create the final sampling clusters. The cell counts were the sums of the weights in each cell. The Dirichlet prior was uniform over all 55,552 cells with a prior sample size selected to insure adequate confidentiality protection. We sampled a complete table from the Dirichlet posterior for each synthetic implicate. An observation was assigned a weight equal to the posterior probability in its final sampling cluster times the civilian non-institutionalized U.S. population as of April 1, 2000 age 18 or older.

The synthetic weight was tested for analytical validity by comparing the pooled results from the 16 synthetic implicates to the analysis from the Gold Standard file of the same earnings and benefit measures that were studied to assess the quality of the final weight itself. When weighted analyses from the synthetic implicates are combined according to the correct formulae, the synthetic weight is just as reliable as the final weight we created for the Gold Standard file. The maximum discrepancy between the weighted Gold Standard analysis and the weight synthetic data analysis is -4.44% and most discrepancies are less than 2% in absolute value.

1.5 Analytical validity testing

Although synthetic data are designed to solve a confidentiality protection problem, the success of this solution is measured by both the degree of protection provided and the user's ability to estimate scientifically interesting quantities reliably. The latter property of the synthetic data is known as analytical (or statistical) validity. Analytical validity exists according to Rubin Rubin (1987) when, at a minimum, estimands can be estimated without bias and their confidence intervals (or the nominal level of significance for hypothesis tests) can be stated accurately. The estimands can be summaries of the univariate distributions of the variables, bivariate measures of association, or multivariate relationships among all variables.

When creating synthetic data, the analyst's goal must be to refrain from imposing prior beliefs about the relationships among the variables. Instead, the synthesizer must be constructed in a manner that allows existing relationships to be expressed with approximately the same degree of precision as they have in the underlying original data. When modeling a particular variable using the Sequential Regression Multivariate Imputation method that is our primary technique, all other variables, powers of these variables and interactions among these variables can potentially be used as explanatory variables even when such a relationship might not seem sensible to a researcher. Of course, due to feasibility constraints, the analyst must choose some subset of variables to go on the right-hand side of the predictive regressions but the goal remains to impose as few prior beliefs as possible. If the analyst is successful in specifying these components of the synthesizer, the result should be analytically valid synthetic data.

Section 6 gives a complete summary of the inference framework and computational formulae for assessing analytical validity. From a theoretical framework, the synthetic data will be analytically valid for the precise set of relations embodied in the posterior predictive distributions used for the synthesis. This theoretical result is reassuring to the

²A PIK is the Census Bureau's internal unique person identifier that replaces Privacy Act protected identifiers, like SSN, on files that have been approved for linking at the individual level.

extent that substantial computational power and flexible methods for estimating complex multivariate distributions can produce reliable posterior predictive distributions. Given the limits of current technology, however, the analytical validity of a particular synthetic data product must be directly assessed. Our method of assessment proceeds as follows. Parallel analyses of a large number of estimands are conducted on the completed Gold Standard data and on the synthetic data. The estimand is averaged over all implicates: four in the case of the completed confidential data and 16 in the case of the synthetic data. Next, the within and between implicate components of the estimand's variance are combined, according to rules that depend upon the precise multiple imputation method used, to generate an estimate of the total variance. The square roots of the diagonal elements of the total variance matrix and the appropriate degrees of freedom are used to form 95% confidence intervals for the completed and synthetic data estimates of the estimand. Ideally, the confidence intervals computed from the synthetic data should cover the confidence intervals computed from the completed data. At a minimum, there should be substantial overlap in the confidence intervals. The point estimates should also be "close," but this result is a by product of confidence interval coverage. In general, the confidence interval in the synthetic data will be wider than the interval computed from the completed confidential data for a specified nominal level, and this loss of precision is part of the cost of confidentiality protection. However, the width of the synthetic confidence interval can be reduced by increasing the number of synthetic implicates. A summary of our analytical validity results follows.

1.5.1 All univariate distributions

We compared the results for univariate distributions of all continuous variables using the first, fifth, tenth, twenty-fifth, fiftieth, seventy-fifth, ninetieth and ninety-fifth percentiles, and the means. Our synthesizer was designed to reproduce univariate distributions through the use of kernel density estimator transformations and inverse transformations. Our comparison of univariate results confirms that the synthesizer worked as designed. Only the wealth-related variables, which have notoriously skewed and multi-modal distributions, proved difficult to synthesize as measured by the univariate distributions. Even the kernel density estimators were not completely successful. One could as reliably compute univariate statistical tables representative of the civilian, non-institutional population age 18 and over on April 1, 2000 from the synthetic data as from the completed data.

We also compare the frequency distributions for categorical variables. These, too, are analytically valid as regards their univariate distributions.

1.5.2 Summary statistics for all workers and for OASDI beneficiaries

Although our synthesizer automatically develops models for subgroups when there are adequate sample sizes, the order in which the subgroups will be formed and tested for sample size adequacy is specified in advance. Consequently, one cannot say *a priori* that results will be analytically valid for all subsamples. We compared the results for all workers and for all OASDI beneficiaries using subsamples constructed on demographic variables and benefit type. This testing focused on important earnings and benefit measures. Work histories, average annual earnings, average indexed monthly earnings (AIME) or average monthly wages (AMW), primary insurance amount (PIA), lifetime earnings, and personal savings account accumulated balances are very similar between the synthetic and completed data files for all major demographic subgroups and all types of benefits. In general, the univariate confidence intervals in the synthetic data cover those in the completed data and are not excessively wide. These results hold whether the reference group is all persons age 18 or older or only OASDI benefit recipients. Overall, the version 4.0 synthetic data have almost complete analytic validity for these tests. This is a notable improvement over all previous versions and, in particular, over version 3.1. See section 6.4.3 for the detailed summary of the results for all workers and section 6.4.2 for the detailed summary of the OASDI recipient results.

1.5.3 Summary statistics by education and foreign born

We also studied summary statistics for several important variables by three-way interactions of race, gender, and education category. This analysis focused on earnings and benefits in 2000. Again, most point estimates were very close and synthetic confidence intervals covered the completed data intervals. In earlier versions, there were problems with certain educational categories. Some problems remain with the groups having no high school diploma or graduate degrees. These problems are usually that the confidence interval in the synthetic data is excessively wide—indicating that the synthesizer had trouble simulating these relationships and reflected a great deal of model uncertainty in the posterior predictive distribution. This is not surprising since these education categories, when cross-classified by race

and gender, contain relatively few individuals in the Gold Standard file. See section 6.4.4 for a detailed summary of these results.

The same analysis was repeated for foreign born individuals, which is a four-way interaction in the underlying data. The results are generally encouraging but in this case the small Gold Standard sample sizes frequently produce very wide (or undefined) synthetic confidence intervals. See section 6.4.5 for a detailed discussion of the foreign born results.

1.5.4 Selected regression model results

We studied the coefficients in selected regression models, fit for the entire sample and for demographic subgroups. Our analysis of the logarithm of total Detailed Earnings Record wage and salary income (deferred and non-deferred at FICA and non-FICA-covered jobs) is representative of earnings analyses. All analyses are markedly improved over version 3.1 of the synthetic data. Most coefficients have some analytic validity—point estimates are similar and synthetic data confidence intervals significantly overlap completed data intervals. There is a detailed discussion of both the successes (most education categories, ethnicity) and the relative failures (actual labor force experience). The earnings analysis is repeated for other earnings measures with similar results.

The analysis of regressions modeling the logarithm of $AIME/AMW$ shows analytical validity for all major demographic groups and virtually all studied variables. The synthetic data can be used to model this variable almost as reliably as the completed data. This is remarkable considering that $AIME/AMW$ was not directly synthesized. Rather, it is derived from the synthetic earnings data.

We also studied regression models for the monthly benefit amount. We believe that the monthly benefit amounts are some of the most accurately synthesized variables and these regression results support this conclusion.

The analytical validity results for measures of earnings and wealth from the SIPP data were less successful. These variables have proven very resistant to synthesis by SRMI and, to date, no adequate alternative technology has been developed.

See section 6.4.8 for a detailed discussion of the regression results.

1.6 Disclosure avoidance assessment

The link of administrative earnings, benefits and SIPP data adds a significant amount of information to an already very detailed survey and could pose potential disclosure risks beyond those originally managed as part of the regular SIPP public use file disclosure avoidance process. The creation of partially synthetic data is meant to prevent a link between these new public use files and the original SIPP public use files, which are already in the public domain. In addition, the synthesis of the earnings data meets the IRS disclosure officer's criteria for properly protecting the federal tax information found in the summary and detailed earnings histories used to create the longitudinal earnings variables in the Gold Standard and public use files. Our disclosure avoidance research uses the principle that a potential intruder would first try to re-identify the source record for a given synthetic data observation in the existing SIPP public use files, which were used to create the SIPP component of our Gold Standard file.

In order to test the effectiveness of the synthetic data in controlling disclosure risk, we conducted two distinct matching exercises between the synthetic data and the Gold Standard. Since the Gold Standard contains actual values of the data items as released in the original SIPP public use files, the Gold Standard variables are the equivalent of the best available information for an intruder attempting to re-identify a record in the synthetic data. Successful matches between the Gold Standard and the synthetic data represent potential disclosure risks. However, for an actual re-identification of any of records that were successfully matched to an existing SIPP public use file, an additional non-trivial step is required—the intruder must make another successful link to exogenous data files that contain direct identifiers such as names, addresses, telephone numbers, *etc.* Hence, the results from our experiments are very conservative estimates of re-identification risk. Nevertheless, we find that the re-identifiable records represent only a very small proportion of the candidate records, less than three percent using the most aggressive technology, and that these correct re-identifications are swamped by a sea of false re-identifications, which a real intruder would not be able to distinguish from the true re-identifications.

The Census Bureau Disclosure Review Board has adopted two standards for disclosure avoidance in partially synthetic data. First, using the best available matching technology, the percentage of true matches relative to the size of the files should not be excessively large. In our case, the true match rate never exceeds three percent of the relevant candidate records. Second, the ratio of true matches to the total number of matches (true and false) should

be close to one-half. We have performed two types of matching exercises, probabilistic and distance-based. The first criterion ensures that very few candidate re-identifications occur. The second criterion ensures that those candidate re-identifications are surrounded by substantial uncertainty as regards their correctness.

We conducted two types of record linking experiments to assess disclosure risk. The first experiment used the Census Bureau's internal probabilistic record linking software to attempt to re-identify the source record of a synthetic file observation in the Gold Standard file. The second experiment used four recently proposed distance-based record linking metrics to attempt the same reidentification. Both experiments were aggressive and conservative.

Aggressive record linking experiments use information that should not be available to a potential intruder but which is available to the analysts conducting the experiment. In our probabilistic record linking, we made aggressive use of the fact that we know the correct linkages between the Gold Standard and synthetic records to estimate the parameters of the agreement score that is used to find candidate matches. In our distance record linking, we made aggressive use of this same knowledge to estimate the full Mahalanobis distance between two records. Such a distance measure uses the covariance structure of the errors in synthesizing the data.

Conservative record linking strategies ensure that the estimated linking rates are upper bounds to what an intruder would calculate. In both experiments, we blocked on the unsynthesized SIPP variables. An intruder would do likewise. To reduce computational burdens, we also segmented the comparison files in a manner that ensured that the true match was always in the segment of the Gold Standard file that was compared to a segment of the synthetic file. Without prior knowledge of the true matches, which would make the record linking exercise superfluous, no intruder could reduce the computational burden with a similar strategy. Because both experiments could always find a correct link in their candidate records, while at the same time the number of at-risk records was artificially limited to reduce computation time, all estimated true match rates are over-estimates.

In the probabilistic record linking experiments, we found true match rates that never exceeded 1.2% overall. The ratio of true to false matches is always around 1/100 and never even approaches unity. In our distance record linking experiments, we found true match rates that never exceed 3%. The ratio of the true to false match rate is not as useful in distance record linking because the false match rate is always one minus the true match rate—every synthetic record has a best match in the Gold Standard file using distance linking techniques. We substituted an analysis of the true match rates based on using the best, second best, and third best distance record linking match candidate. Our analysis shows that there is considerable uncertainty regarding which of these three candidates is the correct match. The ratio of true matches associated with the second or third best candidate to true matches associated with the best candidate hovers around unity.

Both experiments clearly demonstrate that the partially synthetic SIPP/SSA/IRS PUF meets the standards of the Census Disclosure Review Board, which is expected to formally declare the version 4.0 file releaseable before the end of November. Because the public use file is also based on data from SSA and IRS, the consent of their disclosure review officers is also required before the file can be officially released.

1.7 Using the SIPP/SSA/IRS-PUF

This report includes a brief primer on using synthetic data. We explain how to calculate statistical measures on the different synthetic implicate files. Then we explain how to use the control variables placed on those files to properly compute confidence intervals and hypothesis tests. Our primer is not intended to be exhaustive. Rather, it provides a beginning user the wherewithal to process the PUF using standard statistical programming languages like SAS.

1.8 Next steps

Given the length and scope of this project, it is perhaps beneficial at this point to consider what has been accomplished. This collaboration between four government agencies has produced several new data products and advanced the body of knowledge on missing data imputation, assessing the validity of automated statistical modeling, disclosure avoidance techniques, and disclosure risk analysis. In the past six years, we have produced a highly useful compilation of SIPP data that combines five separate panels with edited administrative data from IRS and SSA, a weight to allow meaningful analysis of these combined panels, a set of files that multiply impute all missing data, and a set of synthetic data files that meet disclosure standards of the Census Bureau, the Internal Revenue Service, and the Social Security Administration. For the first time in 30 years, it appears that it will be possible to release lifetime earnings histories taken from administrative records, an accomplishment that will be of enormous benefit to the research community and the general population. This project has been a model for what inter-agency cooperation can accomplish by pooling

the expertise of researchers from the Census Bureau, IRS, SSA, and CBO.

When we began this project, there was a great deal of uncertainty over whether synthesizing techniques could produce micro-data that would preserve relationships among variables and mitigate disclosure risk. In fact, almost none of the enhanced theory or experience with these methods required to complete the project existed. Based on the results at this point, we feel that both these questions can be answered in the affirmative. It is now imperative that outside users be given a chance to test these synthetic data and that the agencies involved develop a system for validating outside results using the Gold Standard in order to promote general confidence in the methods and to permit quality improvements. This process will help us to discover remaining flaws in the synthetic data and improve the synthesizing process, both of which will enable the collaborators to provide useful future updates to this data product, as funding resources permit.

2 Project Background

2.1 Purpose and brief history

In February 2001, a temporary U.S. Treasury Regulation went into effect that allowed the U.S. Census Bureau to obtain administrative W-2 earnings data for certain survey respondents from the Social Security Administration (SSA) and the Internal Revenue Service (IRS) for the purpose of improving core Census Bureau data products. To accomplish the goal of improving the Survey of Income and Program Participation (SIPP), the Census Bureau created an approved project entitled the “Demographic Survey Improvement Project” as a part of the Longitudinal Employer-Household Dynamics (LEHD) Program. Work began on the improvement of the SIPP and on the creation of a new public use file, which is the subject of this report. In February 2003, the temporary Treasury Regulation became final (see *Federal Register*, Vol. 68, No. 13 Tuesday, January 21, 2003, Rules and Regulations, pp. 2691-5).

One of the primary goals of the survey improvement project was to create a new public use file that linked existing SIPP data with the administrative earnings data as well as administrative benefits data maintained by SSA. To this end, a joint committee was created with members from the Census Bureau, the SSA, the IRS, and the Congressional Budget Office (CBO). Individuals with related interests from the staff of the Joint Committee on Taxation (JCT) were also invited to participate. Committee members from the Census Bureau included John Abowd, Nancy Bates, Gary Benedetto, Pat Doyle, Judy Eargle, Sam Hawala, and Martha Stinson, who has served as the coordinator of the project since 2003. SSA has been represented by Susan Grad, Brian Greenberg, Howard James, and Dawn Haines. IRS members included Nick Greenia and Karen Masken. John Sabelhaus participated for the CBO. This committee has guided all major decisions concerning the creation of the public use file.

Beginning with fiscal year 2004, an Inter-Agency Agreement and subsequently a Jointly Financed Cooperative Agreement established an official jointly financed and sponsored project between the Census Bureau and SSA whose main purpose was the research leading to the improvement of the SIPP and the creation of the new public use file. Those agreements provide the basis of the financial and intellectual support for this work. This report summarizes the work done during fiscal year 2006 to finish the creation of the public use file. Inasmuch as the goal is to release a file for use by others outside the development group, this report also includes some history of the project where necessary to understand the final product.

2.2 Overview project description

From the beginning of the project, two over-arching requirements have guided the decisions made by the committee about the type of public use file to create. First, the file should contain micro-data in a format usable by researchers and others familiar with the structure and content of the regular SIPP public use files. Second, the file should stand alone and not be linkable to any of the existing SIPP public use products previously published by the Census Bureau. These criteria led to several other early decisions.

The first major design decision was that the file would contain records for individuals surveyed in one of five SIPP panels, 1990, 1991, 1992, 1993, and 1996, but the panel of origin for each individual would not be revealed. The decision to suppress the panel of origin for the individual was part of the overall confidentiality protection plan for the new PUF. The second major design decision was that the number of variables on the new public use file that came from the SIPP would be limited and would be chosen to facilitate national studies by retirement and disability researchers. The third major design decision was that the primary disclosure avoidance method would be to produce partially synthetic micro-data that could not be re-identified in the existing SIPP public use files. Thus, instead of containing the actual values of SIPP-reported variables, administrative earnings and benefits, the file would contain values that were draws from the joint posterior predictive distribution of the underlying variables conditional on the existing confidential data. The process of synthesizing data is described in detail in section 4.

The committee began its work by selecting the variables to include on the file. The selection process involved detailed discussions between all four agencies and consultations with outside researchers. As part of the process, the Census Bureau created a standardized extract of variables from each SIPP panel and merged these extracts with individual administrative earnings and benefits records. These extracts were combined and named the “Gold Standard” file. (See section 3 for a detailed description of this file.) The Gold Standard file has been revised many times during the past five years as new variables have been added, old ones dropped, and formatting for some variables changed. This file serves as the basis for the creation of the SIPP/SSA/IRS-PUF. It establishes the metadata for each variable, determines the sample of people to be included, and serves as the source data for the modeling required to create the

synthetic data. The Gold Standard file contains data that are Title 13, Title 26, and Title 42 confidential because it commingles Census Bureau, IRS, and SSA data.

The next step in the process was to create a set of synthetic files that replicated the structure of the Gold Standard data. The Census Bureau produced the first such files in late fall 2003, and called it preliminary SIPP/SSA/IRS-PUF version 1.0. Since that time there have been three other preliminary public use files: version 2.0 (fall 2004), version 3.0 (December 2005), and version 3.1 (June 2006). The current preliminary public use file, which is expected to become final, is version 4.0, and is being delivered in conjunction with this report.

After each preliminary public use file was produced, committee members from each agency were responsible for reviewing the file to assess analytical validity and disclosure risk. Analytical validity tests have consisted of comparing univariate distributions, cross-tabulations, moments, and regression coefficients calculated from the synthetic and the completed Gold Standard data. (See section 6 for a detailed discussion of the analytical validity assessment.) Disclosure risk analysis has included probabilistic and distance-based record linking between the synthetic and the Gold Standard files. (See section 7 for a detailed discussion).

3 Creation of the Gold Standard File

The work of creating a new public use SIPP product with linked administrative data began with the creation of a base data set called the “Gold Standard” file. To create this file, we extracted variables from the five SIPP panels conducted in the 1990s (beginning in 1990, 1991, 1992, 1993, 1996, respectively) and merged SSA-provided administrative data from the Summary Earnings Records (SER), Detailed Earnings Records (DER), and the Master Beneficiary Record (MBR). This data compilation serves as the basis for the public use file. We refer to these data as the Gold Standard because they represent the available confidential micro-data that would be used for analysis by an authorized researcher working in a restricted-access facility. Any public use version of these data must, of necessity, closely reproduce the characteristics of the Gold Standard while at the same time taking steps to ensure the confidentiality of the actual data on the sampled individuals.

In this section, we describe each data source, list the variables chosen for inclusion in the Gold Standard file, and explain the major decisions made regarding different types of variables. A complete data dictionary for Gold Standard Version 4.0 accompanies this report. The data dictionary provides exact details about the creation of every variable in the Gold Standard, including the specific source SIPP, SER, DER and MBR variables used.

3.1 SIPP data

We chose the following demographic variables to be included on the Gold Standard file: gender, marital status, race (black/African-American), five categories of education, Hispanic ethnicity, birth date, death date, disability status, number of children, marital history, foreign born indicator, decade arrived in the United States if foreign born, and a spouse identifier that links to the marriage partner if the respondent is married and the spouse was also surveyed. We took the values for these variables at a point in time and thus none of these variables are time-varying in the Gold Standard file. For the time-invariant variables—gender, race, and Hispanic ethnicity, values were taken from the point in the survey when they were first reported, generally wave 1. Values for the other demographic variables were generally chosen from month 8 of the respective SIPP panel (*i.e.*, the last reference month of the second interview). We chose this point because marital, immigration, and disability histories were collected in the wave 2 topical modules and we wanted to take all these variables from the same point in time as nearly as possible. However, if an individual was not surveyed in wave 2 of the SIPP panel either because he or she exited the sample due to attrition from the panel after wave 1 or joined the panel in wave 3 or later, we took values for marital status and the spouse identifier from the closest available point in time. In other words, if marital status was missing in month 8, we checked for a marital status value in months 7, 9, 6, 10, 5, 11, 4, 12, 3, 13, 2, 14, 1, 15, 16, 17, and so on until the end of the panel. We chose the first non-missing value that was found. For individuals whose marital status was taken from a month other than 8, we chose the value for number of children from the same month as marital status. If this value was missing, we did not search in any additional months. For education, we searched over all reported education values in each wave of the SIPP and chose the highest level of education ever reported. Thus gender, marital status, race, education, Hispanic ethnicity, and spouse identifier are never missing in the Gold Standard data because these variables are all reported at some point in the SIPP, and we chose to take self-reported values whenever they were available. Disability status, number of children, marital history, foreign born indicator, and decade arrived in the United States if foreign born are sometimes missing because individuals did not answer the relevant topical modules or because we chose not to search over every available month of SIPP data.

The marital history variables are some of the most complicated historical variables on the Gold Standard file. Most of the information in this history came from the marital history topical module collected in wave 2 of each panel. We supplemented this information by creating a short marital status history that covered the period of the panel from wave 2 forward by using the marital status reported in each month. In the topical module, individuals could report 0, 1, 2, 3, or more than 3 marriages. Dates for the beginning and end of first, second, and most recent marriages were then collected, as well as the reason for a marital termination (death or divorce/separation). If an individual had more than 3 marriages, no dates for those marriages between the second and most recent were collected. We used our short history from the panel period post-wave 2 to check for additional marital events: beginning of a new marriage or ending of an existing marriage. We took account of at most one additional marital history event for an individual. We summarized all this information in a set of 16 variables. They include: *mh_category*, a categorical variable that classifies individuals according to their number of marriages and the type and order of the endings of those marriages; *mh1 – mh7*, a set of flags that provides the same information as *mh_category* but broken down by

event; *flag_mar4t*, an indicator for whether the individual was missing a marriage because the SIPP only collected information on three marriages; age at the time of every reported marital history event.

It is important to understand that the marital history variables may differ from the marital status variable described earlier. In particular if a person reports being married in wave 2, month 8 but is not married at the end of his or her history, this is because a divorce or death occurred over the course of the SIPP panel. Similarly, if a person reports not being married in wave 2, month 8 but is married at the end of his or her history, this is because a marriage occurred during the course of the SIPP panel. Although a person may report only 3 marriages during the topical module, it is possible to have a fourth marriage as part of the marital history because the last marriage occurred during the course of the SIPP panel. However no more than 4 marriages can be recorded in the SIPP from all available sources. There are also some things that must be consistent between the marital history and the wave 2 marital status. If the person reports being divorced or widowed in wave 2, month 8 then at least one divorce or widowhood must occur during that individual's marital history. Likewise if the person reported being married, then the history must contain at least one marriage.

Birth date and death date are unique variables in both their source and treatment. We originally extracted birth date from the first self-reported value in the SIPP survey. However, after several discussions between the Census Bureau and SSA about the measurement error likely to be contained in this variable, we switched to using an administrative birth date. Thus, in the final version of the Gold Standard, we create a variable called *birthdate_pcf* from the Census Personal Characteristics File (PCF), an administrative database that has as one of its inputs the Social Security Numident file. Any individual that has applied for a Social Security Number (SSN) has a record in the Numident file that contains, among other things, SSN and birth date. The administrative record birth date variable (*birthdate_pcf*) serves as the basis for the synthesis process and is comparable to the variable *birthdate* in the synthetic public use files. We chose the administrative source for this important variable in order to insure as much consistency as possible between the administrative earnings and benefits variables and age. Using administrative birth date helps to avoid cases where it appears that people receive retirement benefits prior to age 62, a legal impossibility caused by self-reported birth dates that are several years later than the actual dates.³ We also included a variable called *birthdate_sipp* on the Gold Standard file in order to facilitate re-identification tests that attempt to link the synthetic data back to the Gold Standard. Since the administrative birth date is not an existing SIPP public use variable, anyone attempting to link the new synthetic SIPP/SSA/IRS-PUF to the existing SIPP public use data would have to use the SIPP reported birth date for this purpose.

Death date was extracted from the same PCF file as administrative birth date. In this case, however, the original sources were the Numident file and SSA's Death Master File (DMF), another supplementary file that the Census Bureau receives from SSA that reports deaths every month. The link between SIPP respondents and the PCF was performed using a validated SSN, a process described in more detail in section 3.2.

For economic variables, we included the following annual time series, beginning in 1990 and ending in 1999: weeks worked with pay, weeks worked part-time, total annual hours, family poverty threshold, total family income, total personal income, total personal earnings, welfare program participation and amount of payments, private health/disability program participation and amount of payments, and health insurance coverage and type. Since no individual was followed by a SIPP panel for more than 4 years, these time series arrays contain at least 6 years of missing data for every individual. The exact number and timing of missing years depends upon the original panel. Individuals surveyed in the 1990 panel are missing 1993-1999 data whereas individuals surveyed in the 1996 panel are missing 1990-1995 data. We included the following point-in-time variables: industry and occupation for the main job, chosen from the first available wave, and total net worth, home ownership, home equity, non-housing wealth, and indicators for defined benefit and defined contribution pensions, all taken from topical modules.

Some SIPP variables were purposely omitted from the public use file in order to minimize disclosure risk. Specifically, no data are provided on geography. We include a state of residence variable Gold Standard file but will not release this variable on the public use file. The exact linkage of spouses is the only family relationship data on the file. No other family relationship data are provided on either the Gold Standard or public use files. No panel dating information is provided on the public use file although we retained the panel source variable in the Gold Standard data

³It is worth noting that the administrative birth date is not without some error. Unlike the SIPP reported birthdate, which was edited prior to public release to produce a set of plausible ages, the administrative birthdate contains some values that make individuals in our sample 100 years old or more. However the number of these cases is very limited and we feel that this small error is out-weighed by the general accuracy gains and the benefits to disclosure avoidance.

to facilitate evaluation and testing.

An individual was eligible to be included in the Gold Standard file if he or she met one major requirement: the individual must have been at least 15 years old at the time of the second wave of the SIPP panel in which that person was interviewed. We chose this age because at 15 or older, the SIPP considered the individual to be an adult and asked the full battery of questions. In order to make this determination, we used the variables *popstat* (1990-1993 panels) or *epopstat* (1996 panel) from the wave 2 core data. For those who were not interviewed in wave 2, their age at the end of wave 2 was calculated and if they would have been at least 15, they were kept in sample. It is important to note that these age calculations were done using the self-reported SIPP birth date. We did this in order to reproduce the survey determination of who was eligible to be treated as an adult as accurately as possible in the new public use file.

3.2 IRS/SSA earnings data

Administrative earnings data were extracted from the Master Earnings File (MEF), a historical compilation of earnings reports filed with the IRS by employers (most commonly using the W-2 form). This administrative database is maintained by SSA for the purpose of calculating benefits when workers retire or become disabled. We receive earnings in two forms: Summary Earnings Records (SER) and Detailed Earnings Records (DER). The SER data contain total personal earnings capped at the FICA taxable maximum for each year from 1951-2003. The DER data contain uncapped earnings broken out by employer from 1978-2003. In the Gold Standard file we include the entire annual SER history plus total earnings from 1937 to 1951. From the DER, we create total annual earnings from FICA covered jobs by summing earnings from each employer that was required to withhold social security tax. We also create total annual deferred earnings from FICA covered jobs by summing deferred wages (*i.e.*, contributions to 401(k) plans) from these same employers. We create an analogous set of earnings and deferred earnings variables that pertain to jobs not covered by FICA. Thus, in each year from 1978-2003, the SER earnings variable indicates the amount of FICA covered earnings in a year, up to the taxable maximum, and the set of four DER earnings variables indicates total earnings, uncapped, split between deferred and paid, and FICA and non-FICA covered jobs. The sum of the two DER earnings variables that represent paid wages gives total wages and salary that an individual would report on IRS Form 1040 and which would be taxable under federal income tax laws.

These IRS/SSA earnings variables are matched to the SIPP extracts using a validated Social Security Number. The 1990s SIPP panels collected the SSN from respondents. Using name, address, birth date, gender, and race information, the Census Bureau and SSA validated these self-reports against the SSA Numident file. If the demographic variables collected by the SIPP matched the demographic variables associated with the reported SSN on the Numident, then the SSN was declared valid.⁴ If the demographic variables did not match, an alternative SSN was sought. For individuals who reported that they did not know their SSN, an SSN was sought in the PCF file based on these demographic variables. For individuals who refused to provide an SSN, no match was sought in the PCF and we did not receive earnings records for these individuals. Thus, for individuals without valid SSNs, all the administrative earnings arrays described above were treated as missing data. Approximately 12% of individuals in the gold standard did not have valid SSNs and were, consequently, missing MBR, SER and DER data.

3.3 SSA data

In addition to administrative earnings records, the Census Bureau also received records for SIPP respondents containing information about the type and amount of benefits paid under the Old Age, Survivor, and Disability Insurance Program (OASDI). These SSA data were contained in the Master Beneficiary Record (MBR) file and were linked to the SIPP data using the same method as the earnings data (*i.e.*, validated SSN). The MBR is an extensive and complicated file and, after much deliberation, the decision was made to include only a few variables from it on the Gold Standard file. Specifically, we included the date of initial entitlement to OASDI benefits, the initial reason for receiving these benefits (TOB), and the initial monthly amount paid (MBA). We also included the type and amount of benefit received in April 2000. Using the formulae published by SSA, we calculated average indexed monthly earnings (*AIME*) or average monthly wage (*AMW*) and the primary insurance amount (*PIA*) from the administrative earnings history and included these on the Gold Standard as a help for researchers. However, it is important to note that the *AIME/AMW* and the *PIA* contain no information not already represented in the earnings history. Thus, they can be recreated in the Gold Standard, the completed Gold Standard, or the SIPP/SSA/IRS-PUF data using

⁴Prior to 2003, the process of validating an SSN was performed by a clerical edit using the same information.

alternative assumptions.

3.4 Weight creation and use

One concern that arose early in the process of creating the public use file was the provision of proper weights for a file that pooled SIPP respondents from five separate samples. The 1990-1993 panels contain some overlapping time periods. The official SIPP public use file documentation explains how to pool the published weights for those panels in order to construct a weight that has a well-defined reference population and reference date. However, there is no design guidance for pooling the individuals from all five SIPP panels in order to produce estimates representative of a well-defined target population at a known reference date. In addition, the different SIPP panels over-sample low income individuals and other groups at differential rates. Hence, these survey data can only be used to construct estimates representative of the U.S. civilian non-institutionalized population as of a particular date if an appropriate weight is provided. Thus, another major data activity for this project has been to create an *ex post* weight for the individuals in the Gold Standard file such that each person's weight indicated how many persons in the reference population that SIPP person represented as of a known date. The designated reference population is all individuals age 18 or older in the civilian non-institutionalized U.S. population as of April 1, 2000, the reference date for Census 2000. A full report on the details of this process is provided in section 5.

3.5 Gold Standard data dictionary

The Gold Standard data dictionary for version 4.0 is included as an appendix to this report.

4 Data Completion and Synthesis

4.1 General methodology

In this section, we describe the basic theoretical framework for creating synthetic data. The notation and definitions follow Rubin (1987), which treats multiple imputation of missing data, and Rubin (1993), which is the first paper to define the use of fully synthetic data for confidentiality protection. We adopt enhancements for the application of Sequential Regression Multivariate Imputation (SRMI) to synthetic data from Raghunathan, Reiter and Rubin (2003). We use the formal inference methods for multiple imputation-based partially synthetic data from Little (1993) and Reiter (2003). Finally, we incorporate the formal inference methods for multiple-imputation based partially synthetic data that also have missing data from Reiter (2004). We have attempted to make the notation consistent in this section. Hence, it does not match the original authors' notation.

A finite population contains N entities whose characteristics are known and constitute the f columns of X , $(N \times f)$. A sample of size $n < N$ is drawn from the population. Let the vector I $(N \times 1)$ be defined as $I_i = 1$ if entity i is sampled and $I_i = 0$, otherwise. Data are collected for p variables denoted by the matrix Y $(N \times p)$. Note that the matrix Y is defined for the entire population, not just for the sampled units. Of course, some elements of Y are missing because the entity that constitutes that row was not sampled. Other elements of Y are missing because of item non-response in the sample. (In administrative data, item non-response is equivalent to missing data items on an in-scope administrative record.) Let the matrix R $(N \times p)$ be defined as $R_{ij} = 1$ if the data represented by item Y_{ij} are available in the sample and $R_{ij} = 0$, otherwise. Certain submatrices of Y and R are of interest. Let Y_{inc} $(n \times p)$ be the submatrix of Y that corresponds to the rows for which $I_i = 1$. So Y_{inc} contains the data for all the sampled entities. The complement of Y_{inc} is Y_{exc} , the rows of Y that correspond to the rows for which $I_i = 0$. So Y_{exc} contains the data for all the unsampled entities. Similarly, let R_{obs} $(n \times p)$ be the submatrix of R corresponding to the item missingness for the sampled entities; *i.e.*, those rows for which $I_i = 1$. Finally, define the submatrices Y_{obs} and Y_{mis} as follows

$$Y_{obs,ij} = \begin{cases} Y_{ij}, & \text{if } I_i = 1 \text{ and } R_{ij} = 1 \\ \text{undefined}, & \text{otherwise} \end{cases}$$

and

$$Y_{mis,ij} = \begin{cases} Y_{ij}, & \text{if } I_i = 1 \text{ and } R_{ij} = 0 \\ \text{undefined}, & \text{otherwise} \end{cases}$$

So, the matrix Y_{obs} contains all the sampled values of Y_{ij} that contain data and the matrix Y_{mis} contains all the sampled values of Y_{ij} that are item missing. The observed data are summarized by the set $D = \{X, Y_{obs}, I, R\}$. The following table gives a summary of all these definitions.

General Definitions

N	=	number of individuals in the population
X $(N \times f)$	=	population characteristics of f variables for N individuals, f variables are known for all N individuals in the population
p	=	number of variables for which survey/admin. systems will collect data
Y $(N \times p)$	=	data on p variables for N individuals; only sampled values available
I $(N \times 1)$	=	identifies which individuals from the population were sampled, <i>i.e.</i> tells which rows of Y are non-missing
R $(N \times p)$	=	identifies non-missing elements, <i>i.e.</i> tells which variables are non-missing for which individuals
Y_{obs}	=	observed data, submatrix of Y $(N \times p)$ that contains only elements where individual was sampled and provided data on specific variable
Y_{mis}	=	missing data, submatrix of Y $(N \times p)$ that contains elements where individual was sampled but did not provide data on specific variable
D	=	$\{X, Y_{obs}, I, R\}$ or all known data about individuals in survey sample

In the context of our public use file, the above notation applies as follows: $Y (N \times p)$ is a matrix with one row for every member of the U.S. population age 15 and older at any time between January 1, 1990 and January 1, 1996 and one column for each of p variables that describe these individuals. In our case, there are 173 SIPP variables, 443 SSA/IRS earnings variables, and 5 SSA benefit variables so $p = 621$ and $N = 287$ million. $I (N \times 1)$ contains one row for every member of the U.S. population age 15 and older and $I_i = 1$ when an individual was surveyed by the Census Bureau using the 1990, 1991, 1992, 1993, or 1996 SIPP survey instrument. The matrix I defines $Y_{inc} (n \times p)$ which is a submatrix of $Y (N \times p)$. The I matrix tells which n rows from the population Y matrix were sampled into one of the five SIPP panels and eligible according to age to be in the gold standard: $n = 261,000$. $R(n \times p)$ is a matrix that records which of the n SIPP respondents are missing responses to which of the p variables. $R_{ij} = 1$ if person i has non-missing data for variable j . The R matrix defines $Y_{obs} (n \times p)$ which contains the data we actually observe. The following table provides a summary of these definitions.

Specific Definitions for SIPP/SSA/IRS-PUF

N	=	287 million, i.e. population of U.S.
$X (N \times f)$	=	race, gender, and birth date, known for all individuals on Census short form
p	=	173 SIPP variables, 443 SER/DER variables, 5 SSA benefit variables
$Y (N \times p)$	=	data on all the above variables for entire U.S. population
$I (N \times 1)$	=	identifies which individuals from the population were sampled by the SIPP and included in gold standard
$R (N \times p)$	=	identifies which SIPP and administrative variables are non-missing for which individuals
Y_{obs}	=	observed data, submatrix of $Y (N \times p)$ that contains only elements where individual was sampled by SIPP and data is non-missing
Y_{mis}	=	missing data, submatrix of $Y (N \times p)$ that contains elements where individual was sampled by SIPP but did not provide data on specific variable
D	=	$\{X, Y_{obs}, I, R\}$ or all known data about individuals in the SIPP samples

In the classic Rubin (1987) missing data application, Y_{mis} is imputed m times by sampling from $p(Y_{mis} | D)$, the posterior predictive distribution of Y_{mis} given D . The completed data consist of m sets $D^{(\ell)} = \{D, Y_{mis}^{(\ell)}\}$, where $Y_{mis}^{(\ell)}$ is the ℓ^{th} draw from $p(Y_{mis} | D)$ and is called the ℓ^{th} implicate. The basic insight for using synthetic data as part of a confidentiality protection system is that sampled individuals can be treated as having missing data for some or all variables even if they provided valid data. When these data are “completed” in the same manner as described above, namely by drawing from the posterior predictive distribution of Y_{mis} , $p(Y_{mis} | D)$, a file is produced that remains statistically valid but no longer contains the sampled individuals values for the variables that were synthesized.

In our application of synthetic data methods to the linked SIPP and administrative data, we first use Rubin’s general multiple imputation method to complete our missing data. Next, we used this same method to create synthetic data. It is important to note that data resulting from this process are most accurately described as “partially” synthetic data. The terms “partially synthetic” and “fully synthetic” are now used in the statistics literature to distinguish between two related synthetic data generating models. Partially synthetic data are created using an actual sample of the population (*i.e.* the actual SIPP surveys) as source records so that a record in the partially synthetic data is based upon an actual record from the underlying survey. Fully synthetic data are created by sampling from a synthetic population in which the unsampled entities from the original survey have synthetic values for all variables from the survey. Fully synthetic samples are created by using all the known population characteristic variables to generating synthetic values for all survey variables conditional on the known population characteristics. Thus fully synthetic implicates do not have an actual source record in the original survey and can be described as fictitious entities. This project did not attempt to create fully synthetic data.

The major focus of the synthesizing process is to obtain a good estimate of the posterior predictive distribution

(PPD) for all the variables to be completed and synthesized. We now discuss the computational formulae for estimation and sampling from the PPD. More general methods exist, such as Markov Chain Monte Carlo, but the methods summarized herein are the ones used by this particular project.

To begin, an explicit representation of D is required. As defined above $D = \{X, Y_{obs}, I, R\}$. While, in principle, the analyst at the Census Bureau has access to X , the population characteristics, in the applications described in this section, only the rows of X corresponding to $I_i = 1$ are used.⁵ Hence, there is no practical difference between X and Y_{obs} for our synthetic data modeling. Complete data are guaranteed for X but nevertheless many variables in X require confidentiality protection before they can be placed in a public use data file. In this section, we adopt the notational convention that a variable appears in X if it is always available when $I_i = 1$ and it never requires confidentiality protection. Otherwise, the variable is included in Y_{obs} . This set of X variables can be empty without affecting the discussion below.

We describe two methods: Bayesian bootstrap (BB) and SRMI.⁶ In both of these methods, we apply the principle of estimating the conditional distribution of group of variables (columns of Y) conditional on all other columns. For each distinct group of variables in Y , the columns of D are partitioned into four mutually exclusive sets: grouping variables, conditioning variables, dependent variables, and ignored variables. Grouping variables are used to stratify D such that a separate PPD is estimated in each stratum. Conditioning variables are a list of potential right-hand-side variables to be entered linearly in model-based estimation of the PPD. Dependent variables are those for which the PPD is being estimated. Finally, ignored variables are all other columns of (X, Y_{obs}) . For purposes of doing the computations below, the data matrix (X, Y_{obs}) should be interpreted as including any variables that have been calculated as exact functions of the available data. Hence, the dimensionality of the matrices used below potentially exceeds $f + p$.

4.2 Bayesian Bootstrap

The Bayesian bootstrap was originally defined by Rubin (1981). As explained therein, the BB is used to simulate the posterior distribution of the parameter whereas the regular bootstrap simulates the sampling distribution of the parameter. Whereas a conventional bootstrap assumes that the sample CDF is equal to the population CDF, the BB properly accounts for the uncertainty of the sample CDF.

4.2.1 Generic BB algorithm

The notation used to describe the BB algorithm in this subsection is generic and does not refer to the matrices defined elsewhere. Let X ($n \times k$) be the source data matrix and Y ($s \times k$) be the target data matrix. This means that we want to construct an $s \times k$ Bayesian bootstrap sample from an $n \times k$ matrix of source data. Each BB replicate ℓ is a unique $Y^{(\ell)}$.

1. Draw $n - 1$ random variables from $U(0, 1)$.
2. Sort u_i ascending and let $u_{(i)}$ denote the order statistics from lowest to highest. Define $u_{(0)} = 0$ and $u_{(n)} = 1$.
3. For $i = 1, \dots, n$, let $\hat{p}_i = u_{(i)} - u_{(i-1)}$.
4. For $j = 1, \dots, s$ sample with replacement from the rows X using \hat{p}_i as the probability of selecting row i . Place the sampled row into Y_j .
5. Repeat from step 1 for as many BB replicates as desired.

In other words, beginning with a data matrix, X , that contains values for the k variables of interest, this process assigns a probability of choosing a given observation from X to provide data to a corresponding observation in Y for the k variables. The set of probabilities constitutes a non-parametric representation of the posterior distribution from which the sampling is done. In a conventional bootstrap, because of the assumption that the sample CDF is equivalent to the population CDF, each observation in X would be assigned probability $\frac{1}{n}$ of being chosen. There would be no

⁵An exception is the process used to create and synthesize the *ex post* weight, which is described in section 5. In that process the full matrix X was used.

⁶For a description of the Bayesian bootstrap see Rubin (1981). For a description of SRMI in its original application to missing data problems see Raghunathan et al. (1998).

uncertainty in what probability would be assigned to a given observation. However, the Bayesian bootstrap accounts for the fact that the sample CDF is not the population CDF and hence does not assign equal probability to each observation. To better understand this concept, consider the example of $k = 1$ where the variable of interest, x_1 , is an indicator variable. Suppose that for 75% of the sample of individuals, $x_1 = 1$ and that $x_1 = 0$ for the remaining 25%. In a conventional bootstrap, with each individual assigned a probability of $\frac{1}{n}$ of being chosen, the CDF used for sampling would always give $x_1 = 1$ a 0.75 probability and $x_1 = 0$ a 0.25 probability. The resulting target matrix Y would not necessarily have a realized 75%/25% frequency distribution for the two values for x_1 but all the bootstrap samples would have been drawn from such a distribution. In a Bayesian bootstrap, when each source record is assigned a unique probability whose expected value is $\frac{1}{n}$, the CDF used for BB sampling might have 73% versus 27% probability of drawing $x_1 = 1$ or 0. The next BB might have 76% versus 24%. The variation in the BB probabilities reflects the fact that the sample proportion of 75% in X is an estimate of the probability that $x_1 = 1$.

4.2.2 Bayesian bootstrap application

Choose grouping variables such that the rows of (X, Y_{obs}) can be assumed to come from the same joint distribution within each group defined by the unique combinations of values of the grouping variables. Some collapsing of categories may be required and is described later under implementation details. What is required is the creation of G groups based on the values of the variables in the grouping variable list. It is essential to the success of the Bayesian bootstrap in accurately replicating statistical properties of the data that the observations in a given source (donor) group and a given target (donee) group be as homogenous as possible. Thus, ideally, a large list of grouping variables should be chosen initially. One of the main advantages of the Bayesian bootstrap is that the group sizes do not have to be as large as groups where parametric modeling is done. Another advantage that is described below is that groups of dependent variables can be done together. This method also helps to preserve the statistical properties of the data by keeping intact relationships among variables.

In the BB application, none of the grouping variables can contain missing data. There are no conditioning variables because no linear model is used. The dependent variables consist of all columns j of Y for which $R_{ij} = 0$ for some i . The ignored variable list consists of all variables that are neither grouping variables nor dependent variables. We first describe the application of BB to the missing data problem. This is complicated if the missing data pattern is non-monotone as defined in Rubin Rubin (1987). For the moment, assume that the missing data pattern is monotone. Then, proceed through the dependent variables in groups constructed as follows:

1. All dependent variables with missing data exactly comparable to the variable with the least missing data; *i.e.*, all j for which $R_{ij} = 0$ if and only if $R_{ij^*} = 0$, where j^* is the column index of the variable with the least missing data. This is dependent variable group 1.
2. Remove all variables from the dependent variable list that are already in a group. Let j^* represent the column index of the variable with the least missing data from among those dependent variables that remain. Group all dependent variables with missing data exactly comparable to the variable indexed by j^* ; *i.e.*, all j for which $R_{ij} = 0$ if and only if $R_{ij^*} = 0$. This is dependent variable group h .
3. Increment h and repeat step 2 until no dependent variables remain.

This defines H dependent variable groups. Initialize the BB missing data algorithm by placing all dependent variables into the ignored variable list and setting $h = 1$.

1. Remove the variables in group h from the ignored variable list and place them in the dependent variable list.
2. For $g = 1, \dots, G$, BB the rows of Y_{mis} (target data matrix) using the rows of Y_{obs} as the source data matrix. Repeat the BB m times to get m imputations $Y_{mis}^{(\ell)}$.
3. Put the dependent variables in group h back into the list of ignored variables.
4. If $h < H$ then increment h and return to step 1; otherwise, stop.

The result is m completed data sets. When the missing data are not monotone, the BB algorithm can be used to get starting values for other algorithms described below, in particular, SRMI. The BB algorithm can also be used for synthesizing data. In this case, simply treat all observations as missing and use the above steps to find donors for every individual in the data.

4.3 Sequential Regression Multivariate Imputation

Sequential Regression Multivariate Imputation (SRMI) was first proposed as a general technique for multiple imputation of missing data by Raghunathan et al. (1998). Raghunathan, Reiter and Rubin (2003) extend the method to confidentiality protection. Abowd and Woodcock (2001) use the SRMI method for confidentiality protection combined with missing data imputation. Although the formulae for SRMI can be stated generically using joint probability distributions like $p(Y_{mis}|D)$, almost all applications assume that the entities that constitute the rows of (X, Y_{inc}) have been sampled independently. Nothing in the generic statement of the problem prohibits dependent sampling; however, as a practical matter, formalizing this dependence while implementing SRMI is complicated. Abowd and Woodcock (2001) illustrate these complications for the case of longitudinally linked employer-employee data. The algorithms are summarized below ignoring the complications associated with dependent sampling.

4.3.1 Definitions and general algorithm

In SRMI, the analyst cycles iteratively through the dependent variable list. In any given iteration, conditioning data may be taken from either the current or the previous iteration depending upon the location of the current dependent variable in the variable list. For missing data applications, the procedure is normally iterated until the effect of this conditioning has been minimized. In synthetic data applications, the conditioning values are the same regardless of the position of the variable in the dependent variable list and so iteration is not required.⁷

Let Y_j denote the current dependent variable and let $Y_{\sim j}$ denote all other columns of Y . The general algorithm is most cleanly stated for the missing data case. The refinements for the partially synthetic data case will be noted below.

For each dependent variable, the analyst selects grouping variables, conditioning variables and ignored variables. The grouping variables stratify the estimation into G mutually exclusive and exhaustive groups as illustrated in section (4.2.2). The conditioning variables may include all columns of $(X, Y_{\sim j})$, including columns that are created to allow for nonlinearities in the conditional relations. The ignored variables are all columns of $(X, Y_{\sim j})$ not included among the conditioning variables. We wish to generate m implicates $Y_{mis}^{(\ell)}$. SRMI is an iterative procedure. Denote the interim values of implicate ℓ as $Y_{mis}^{(\ell, s)}$. Initialize $\ell = 1$ and $s = 1$. Initialize $Y_{mis}^{(1, 0)}$ using Bayesian bootstrap methods.

1. For $j = 1, \dots, p$:

(a) If $\ell = 1$ then estimate

$$p\left(Y_j|X, Y_{obs, \sim j}, Y_{mis, 1}^{(\ell, s-1)}, \dots, Y_{mis, j-1}^{(\ell, s-1)}, Y_{mis, j+1}^{(\ell, s)}, \dots, Y_{mis, p}^{(\ell, s)}\right)$$

(b) Fill $Y_{mis, j}^{(\ell, s)}$ with data sampled from

$$p\left(Y_j|X, Y_{obs, \sim j}, Y_{mis, 1}^{(\ell, s-1)}, \dots, Y_{mis, j-1}^{(\ell, s-1)}, Y_{mis, j+1}^{(\ell, s)}, \dots, Y_{mis, p}^{(\ell, s)}\right)$$

2. If converged then

(a) Set $Y_{mis}^{(\ell)} = Y_{mis}^{(\ell, s)}$.

(b) Increment ℓ .

(c) Reinitialize $Y_{mis}^{(\ell, 0)} = Y_{mis}^{(\ell-1, s)}$

(d) Reinitialize $s = 1$

⁷An exception to this statement occurs when the data to be synthesized have exact logical dependencies among the variables. In this case a parent/child tree is used to coordinate these dependencies. The conditioning data for a particular variable will include the results of the synthesis of variables that were antecedents in the parent/child tree (parents). Iterating this process, however, simply produces another synthetic implicate.

3. If $\ell \leq m$, go to 1.

The test for convergence is not formal. In practice s is often limited to 10 or less. The algorithm estimates the joint distribution $p(Y_{mis}|D)$ by iterating over each conditional distribution $p(Y_{mis,j}|D)$ and filling the “data matrix” with imputed values based on the previous iteration’s estimate of $p(Y_{mis}|D)$. Once the estimation has converged, the implicates are all drawn from the same estimate of $p(Y_{mis}|D)$. However, the completion of D for each implicate results in different conditioning data for the draws. In the implementation of the algorithm, one cycles over the grouping variables $g = 1, \dots, G$ performing the entire algorithm for each homogeneous group. In steps 1.a and 1.b only the conditioning variables appropriate for $Y_{mis,j}$ in conditioning group g are actually included in the conditioning set. The initial selection of these variables is dependent on the analyst. However after the variables are tentatively included in the model the Bayes Information Criterion (BIC) is used to reduce the variable list by eliminating variables that have a posterior odds ratio below a pre-specified level. The posterior odds ratio cutoff for the BIC value in this variable selection mechanism can be controlled by the analyst. See Abowd and Woodcock (2001) for details.

4.4 Summary of synthetic data production

We now provide specific details about the process used to create synthetic data for this project. The first step of the process was to multiply impute all missing data. Missing data in our sample are due to survey item non-response and to out-of-scope survey years. Failing to provide an answer to the question about whether an individual was born in the United States or a foreign country is an example of item non-response. Missing income in 1996 because the individual was surveyed in the 1990 SIPP panel, which ended before 1996, is an example of missing due to out-of-scope survey years. The goal of the first step is to impute values for every variable whenever it is missing due to item non-response or out-of-scope survey years. We call this “completing the data,” because the result of this first step is a set of files that contain all the original data plus imputed values when the original data were missing. Each one of these files is then referred to as a “completed” data set.

Regular missing data, which we multiply impute, result from item non-response or an out-of-scope survey year. Structurally missing data occur when an item is missing due to the logical structure of a set of variables in the survey or administrative record. Structurally missing data still exist in our completed and synthetic data—every individual will not necessarily have a value for every variable. For example, an individual who was born in the United States will have structurally missing data for the variable that indicates which decade the person immigrated to the United States. For survey data, structurally missing values occur when the skip logic of the survey dictates that a question is not appropriate because of the response given to a prior question. Administrative record data have a similar, albeit implicit, structure. Statisticians usually call such values “structural zeroes.” Structurally missing data are never completed (*i.e.*, imputed) because they do not represent missing information. In contrast, regular missing data, which we complete by multiple imputation, do represent a failure on the part of the survey or administrative records to capture certain information. In this report we use the term “missing” to mean “missing-to-be-completed” and will explicitly describe any other data that are missing as “structurally missing.”

Completing data involves choosing a model for each variable with missing data. We used the SRMI methodology to impute missing values for most of the SIPP variables. The few exceptions are described in 4.5.5 where we give details about the modeling for each variable. We used the BB technique to handle missing data due to missing SSNs. When an individual failed to provide an SSN that could be validated, we could not link that individual to the administrative databases (PCF, SER, DER, and SSA benefits) and, as a result, several hundred administrative variables were missing. One approach to this problem would have been to use the SRMI methodology to model each individual administrative variable and impute missing values. However the magnitude of this task and concern about the need to preserve internal consistency among all the administrative variables, led us to choose the BB completion method for the SSN variable. This method allowed us to choose an appropriate donor record with a non-missing SSN which provided the complete set of administrative variables: PCF (birth date, death date), SER and DER (earnings), and MBR (SSA benefits). Once the SSN had been completed, we treated all administrative data as completed. If a validated SSN did not have a record in a particular administrative database, we treated these data as structurally missing. In other words, no Master Beneficiary Record meant the person had not received benefits from SSA under a program that would generate an MBR record and no DER job records meant the person had not earned federally taxable income since 1978. Once again, in the completed administrative data, an individual does not have a value for every variable. But individuals who were originally missing SSNs now have donated SSNs which link to administrative data that is

either present or structurally missing.

One important feature of how we applied the BB is worth mentioning. When both members of a married couple were both missing SSNs, we chose a donor couple based on couple characteristics instead of two separate individual donors. In this way we hoped to preserve the important effects of marriage on SSA benefits. When only one member of a couple was missing an SSN, we also chose a donor couple based on couple characteristics but then only used the donated SSN for the couple member with the missing SSN. By using this method, we were able to choose a donor couple that resembled the couple with the single missing SSN and a donor spouse who looked like the donee and was married to someone who looked like the donee's spouse.

The actual process of completing data is iterative. We begin with a base data set that contains only original, non-missing data. We then use the BB to complete the SSN and hence the administrative data. Donors are chosen on the basis of non-missing SIPP variables. This data set serves as the input for the SIPP data completion stage using SRMI. Models are estimated using originally non-missing dependent variables and any available non-missing explanatory variables from either the administrative or SIPP data. Variables are modeled beginning with the variable with the least missing data and progressing to the variable with the most missing data. As models are estimated and missing values are imputed, the data set is updated to include the imputed values. Hence, for the first variable modeled, almost all other SIPP variables will have missing values and hence a number of cases will be excluded from the estimation in this first round. As variables are completed and the data set is updated, there will be fewer and fewer missing values, and increasingly more cases available for model estimation. The end product of the SRMI process is a data set that contains completed administrative and SIPP variables.

We then iterate the process. We perform the Bayesian bootstrap again to complete the SSN, this time using the updated, completed SIPP variables from the end product of iteration 1. We then use SRMI to estimate models for the SIPP variables again. As in the first iteration, only originally non-missing dependent variables are used in model estimation. However beginning with the second iteration, the first variable to be modeled uses explanatory variables from the completed data that was the output of iteration 1. This prevents the exclusion of any cases due to missing data. The second variable to be modeled uses the most up-to-date values for variable 1, *i.e.*, the values imputed in iteration 2, and the completed data from iteration 1 for every other variable. The sequential estimation progresses until the last variable, which uses imputed values from iteration 2 for all explanatory variables. In this manner, the modeling is always done with the most up-to-date imputed values available, allowing the modeling to improve itself over iterations. At the conclusion of this second SRMI step, another completed data set is generated which has updated values for all the SIPP and administrative variables.

As part of the creation of version 3.0 of the preliminary public use file, we performed 8 iterations of missing data completion as described. As part of the creation of version 4.0 of the preliminary public use file, we performed one additional iteration of missing data completion. This was done for two reasons. First, our experience modeling variables over the past year led us to make many improvements that we wished to implement both in the data completion and data synthesis phases. Second, Yves Thibaudeau, from the Census Bureau Statistical Research Division (SRD), provided us with new 1996 SIPP data for home equity. These data were the result of an on-going research project at SRD, sponsored by SSA, to improve the imputation models for some of the variables collected in the wealth topical module in wave 3 of the 1996 panel. Our hope was that these improved starting data would lead to better models for our completion and synthesis of the wealth variables.

The SRMI method estimates the posterior distribution of the regression parameters (coefficients and variance of the error) and draws from this distribution to obtain parameters used to impute values. We impute multiple times, meaning we take multiple draws from the posterior distribution of the regression parameters. The data product that results is actually a set of files called the completed data implicates. Each implicate has an identical structure (same number of observations, variables, *etc.*) and contains identical data in cases where the information was originally non-missing. For example, if total net worth was non-missing for 75% of the individuals in the sample, then 75% of the observations in each implicate file would have identical values for total net worth. The remaining 25% of the observations would have different values of total net worth across implicate files because of the multiple imputation. The implicates are generated by 4 separate SRMI processes, which is necessary because of the inter-related nature of the variables. Once a variable has been completed, its updated value is used as a right-hand-side variable in the imputation process for other variables. For example, once total net worth has been completed, its updated value will be used to impute a missing value for total income in 1990. Thus, in order to maintain internal consistency within an

implicate file, each implicate must be generated separately. For version 4.0, we created four missing data implicates.

Because of the many iterations necessary to complete the data, the majority of the computing time spent creating a synthetic data set is actually spent dealing with missing values. Once the data are completed and contain no missing data except for structurally missing items, the final step of actually synthesizing all the data is takes much less time (*i.e.*, several weeks versus several months). Synthetic implicates are just like completed data implicates except that every individual has his or her values imputed, variable by variable, conditional on the completed data.⁸ For example, in the case of total net worth, in the data completion phase 25% of individuals received imputed values to replace originally missing data. In the synthesizing phase, 100% of individuals received an imputed value to replace their original data, whether it was missing or not. Synthesizing data is in essence like doing one more iteration of missing data completion except everyone's data has to be completed.

The completed data from the appropriate 9th iteration implicate serve as the input for estimating the PPD used in the synthesizing phase. SRMI models are estimated using only originally non-missing dependent variables and completed explanatory variables. Explanatory variables thus contain either original non-missing data or imputed values from the 9th iteration.⁹ We take a draw from the distribution of regression parameters and then impute a value based upon the most up-to-date synthetic data. This means that while the synthetic variables are not used in the model estimation, they are used to impute other synthetic values. For example, when estimating a model for total income in 1990, the values of total net worth used as explanatory variables would come from the 9th iteration completed data. However, when taking draws from the posterior predictive distribution for total income in 1990 in order to generate the synthetic total income 1990 variable, the synthetic value of total net worth would be used if this variable had been previously synthesized. Otherwise, the value from the 9th iteration of completed data was used.

Each one of the completed data implicates serves as the basis for creating four synthetic implicates. Since there are four completed data implicates, there are four separate input files to the synthesizing process. Each completed data implicate then has four distinct modeling steps and produces four separate draws for the regression parameters and four separate sets of synthesized values. This procedure preserves the internal consistency of each implicate file. In the end there are 16 synthetic implicates.

4.5 Modeling details

The actual implementation of either a Bayesian bootstrap iteration or an SRMI iteration is controlled by a SAS program that contains information about every variable and, based on this information, executes the appropriate modeling routines. The critical information that the analyst must provide for every variable is variable type, parent-child relationships, restrictions, level, and a set of grouping and conditioning variables to use in modeling. In this section we define these terms and explain how we assigned values in general. In the next section we list the specific values chosen for every variable.

4.5.1 Types of variables

The first information the analyst must provide about a variable to be completed and synthesized is the variable type. There are three major types of variables in the public use file: continuous, binary discrete, and categorical. The variable type determines which estimation routine will be used for the modeling step. We describe each in turn.

For continuous variables, the imputation model is a normal linear regression, which means that the posterior predictive distribution is based on the normal/inverted gamma posterior distribution for the parameters of a normal linear regression. Under an appropriate uninformative or conjugate prior, the posterior predictive distribution for the variable under study is normal (given the conditioning variables and the standard error of the equation). If the univariate

⁸Reiter (2004) distinguishes between the models used for the missing data imputation and those used for the synthesis, indicating that these models should not be the same if different conditions apply to the selection of values to be synthesized as compared to those that are missing. We fully implemented this distinction. Estimation and sampling from the posterior predictive distribution correctly reflects differences in the conditioning information. For example, to sample a synthetic birth date, we first estimated the PPD for birth date unconditional on range restrictions. When we sampled from the birth data PPD, we imposed the range restrictions discussed below using accept/reject resampling from the unconditional PPD.

⁹This final step of model estimation in the synthesizing phase is in essence a repeat of the estimation done in the 9th iteration of missing data completion. This is because there is no updating of the explanatory variables. The explanatory variables always come from the completed data set that was generated in the 8th iteration of data completion. In fact if we had stored the parameter distribution results from the last round of data completion, we could skip this final model estimation step altogether and use the model results from the data completion phase. However, our programs are not set up to operate in this manner so this has been left for future research.

distribution of the variable we are trying to synthesize, y_k , differs greatly from conditional normality, the distribution of the synthetic values will differ from that of the confidential values. To handle this situation, we transform the confidential data so that they have an approximately normal distribution, estimate the posterior predictive model on the transformed data, and perform the inverse transformation on the imputed values.

The first step is to obtain an estimate of the unconditional distribution of y_k . Since the exact parametric distribution of y_k is unknown, we use a nonparametric estimator; namely, the kernel density estimate \hat{K} . For technical reasons, the kernel density estimator (KDE) is computed from a Bayesian bootstrap sample of y_k , not the exact Gold Standard copy of y_k . The KDE \hat{K} is estimated separately for each set of grouping variable values. In addition, for each set of grouping variable values, the transformation is also estimated and applied to other continuous conditioning variables when appropriate (*e.g.*, if y_k is DER earnings this year and one of the conditioning variables is DER earnings next year, then both variables are transformed by an appropriate KDE estimate of each of their univariate distributions). Next we use the estimated KDEs to transform the actual dependent variable and any appropriate independent variables to normality. For each observation y_k , obtain the transformed value $y'_k = \Phi^{-1}(\hat{K}(y_k))$, where Φ denotes the standard normal CDF. By construction, the y'_k have a standard normal distribution. Next, estimate the regression of y'_k on its (possibly transformed) predictor variables to get an estimate of the posterior predictive distribution of y'_k . Sample synthetic values \tilde{y}'_k from this posterior predictive distribution. The synthetic values are normally distributed with conditional mean and variance defined by the regression model.¹⁰ After standardizing the \tilde{y}'_k to have zero mean and unit variance, compute the inverse transformation $\tilde{y}_k = \hat{K}^{-1}(\Phi(\tilde{y}'_k))$. The imputed values \tilde{y}_k are distributed according to \hat{K} , preserving the univariate distribution of the underlying confidential data. Further details of this procedure can be found in Woodcock and Benedetto (2006).

For binary discrete variables, the PPD is based on the asymptotic posterior distribution of the parameters of a logistic regression model. As described in section 4.3, we first split our sample of SIPP respondents into homogenous sub-groups using a set of grouping variables (sometimes called by-variables because they specify the subsets of observations that will be used for a particular model). Next, we estimated logistic regression models for each sub-group. We encountered problems with this approach when some sub-groups did not have enough variation to make the computation of a unique maximum likelihood estimate feasible. In other words, for some sub-groups, there were some combinations of right-hand-side variables that perfectly determined some value of the dependent variable. This problem, which is well known in the logistic regression literature see Albert and Anderson (1984), created a continuum of maximizers and prevented convergence in the algorithm used to maximize the likelihood function.¹¹ Because of this problem, known as quasi-separation, the results of the logistic regressions were sometimes unreliable and the coefficients had very large standard errors. The problem of partial ordering of the dependent variable in a logistic regression, which causes the log likelihood function not to have an interior maximum even though it is globally concave, is usually handled by respecifying the logistic regression. Failure to do so causes numerous problems with our synthesizer—in particular, the BIC-based automatic variable selection drops too many variables, if not all of them, and the draws from the posterior predictive distribution are extremely dispersed.

We believe that we used well-formulated logistical regression models and that none of the conditioning variables (sometimes called x-variables because they serve as right-hand side variables in the statistical models) had structurally determined relationships with the dependent variable. Hence, we believe that the quasi-complete separation problem was actually a sample size issue. Some of the sub-groups were simply too small. If we were to have large enough sub-group samples, every combination of x-variables and responses would eventually take on some positive probability for every sub-group. That is, we believe that the problem was sampling zeroes, not structural zeroes. Hence, we addressed this problem by using an informative prior on the logistic regression probabilities that is implemented using data augmentation; see Tanner (1996). The augmenting data matrix consists of one record for each potential combination of discrete conditioning variables and each discrete outcome. This imposes an informative Dirichlet prior on the space of outcomes of the logistic regression. The augmenting data provide the variation guaranteed to create a unique estimator for the posterior mode (equivalent to the maximum likelihood estimator in this case). However,

¹⁰This explanation is simplified. We take proper account of the inverted-gamma distribution on the standard error of the regression. Our procedure samples from the posterior distribution of the standard error of the regression, conditional on the sample error sum of squares and degrees of freedom. The sampled value of the regression equation standard error is used in the conditional normal posterior distribution of the regression coefficients.

¹¹In the SAS logistic regression procedure, this error is reported as the warning for possible “QUASI-COMPLETE SEPARATION OF DATA POINTS.”

the effects of the informative prior are dominated by the original data matrix in determining the parameter estimates except when one of the sampling zeroes occurs in a particular sub-group. Then, the prior distribution ensures a unique posterior mode.

For categorical variables, the PPD is based on the asymptotic posterior distribution of the parameters of a binary tree of logistic regression models that are used to model each level of the categorical variable successively as branches in the binary tree. The categorical variable modeling program looks for an equal split of individuals across categories, thus lumping some of the original categories together, and then models the probability that a person falls in either the first group or the second group. Then, within these two groups, another split is done and the probability that a person falls into one or the other of these subcategories is modeled. The binary tree continues until all the original categories have been modeled. Finally, the binary tree is synthesized and the synthetic values are used to recreate a synthetic value of the original categorical variable. For example, when the industry variable with four categories is modeled, the program might first split people into groups based on those with $ind_4cat = \{1, 2\}$ and those with $ind_4cat = \{3, 4\}$. It will then split the groups again in order to model $ind_4cat = 3$ versus 4 and $ind_4cat = 1$ versus 2. After the modeling is finished, a new synthesized ind_4cat variable is created that takes on values 1 to 4.

4.5.2 Parent-child relationships and constrained variables

Next the analyst must provide information that appropriately accounts for explicit relationships among the original variables that need to be preserved in the synthetic data. We have developed two tools for handling these relationships.

Our first tool is to specify parent-child relationships. We define parent variables as those that restrict which observations of another variable are present and which observations are structurally missing. These parent-child relations formalize the skip patterns in the SIPP survey instrument and the logical dependencies in the administrative records. A parent variable determines the universe of observations that are in scope to estimate the model for the associated child variable and will receive an imputed value following the estimation. If the parent variable indicates that the child variable is structurally missing (out of the universe) for an individual, then this observation will not be included in the estimation nor will it receive an imputed value. Instead, it will be set to SAS missing. An example of this type of relationship can be constructed from the variables *foreign_born* and *time_arrive_usa*. *Foreign_born* is the parent variable and takes a value of zero or one for everyone in the data set. It controls whether an individual is in scope to have a value for *time_arrive_usa*, the child variable. If a person was born outside the US, then that person should have a value for decade of arrival in the U.S. This value may be originally missing or not, but when $foreign_born = 1$, the person is in scope to contribute data to the estimation of the model for *time_arrive_usa* and will receive an imputed value for this variable that either replaces the missing data or synthesizes the original data. In this manner, we can prevent structurally missing data from skewing our modeling and we can also ensure that only the appropriate people receive a value for *time_arrive_usa*. In this example, the child variable is in-scope only when the parent variable takes a specific value ($foreign_born = 1$). However, the method generalizes so the parent can take on a range of values. For instance, a person is in-scope to have a value for weeks worked part-time if weeks worked with pay is greater than or equal to one and less than or equal to five. In other words, as long as weeks worked with pay is positive, the person is in-scope to have a value for weeks worked part-time. If a person works a full month but never part-time, that person will have weeks worked with pay equal to four or five (depending on the month) and weeks worked part-time equal to zero. If a person does not work at all in a month, that person will have weeks worked with pay equal to zero and weeks worked part-time will be SAS missing.

Our second tool for handling relationships among variables is to place restrictions or constraints on some variables. Constraints do not restrict which observations are used in estimation nor do they restrict which observations receive an imputed or synthetic value. Instead, constraints specify a minimum and maximum value that restricts the range of draws from the posterior predictive distribution. For example, we synthesize birth date for every individual regardless of the value of any other variables. Thus, there is no parent variable for birth date. However the synthesized value for birth date must be consistent with the age requirements for any SSA benefits received by an individual. For example, if the individual began receiving retirement benefits in 1980, he or she must have been born by 1918 at the latest in order to be at least 62 years old by the time initial retirement benefit receipt. Thus, restrictions are imposed on birth date by the initial type of benefit and date of initial entitlement variables. Our programs impose these constraints by calculating what we term “utility variables” that contain these maximum and minimum values for every constrained variable. When we draw from the posterior predictive distribution for a constrained variable, the candidate sampled value is compared to the maximum and minimum for this individual and if the candidate draw falls

outside the specified range, another draw is taken. This comparison and re-sampling is repeated until the candidate sampled value satisfies the constraints or 100 candidate draws have been performed—at which point the value is set equal to the closest boundary (*i.e.*, if the value is over the maximum on the 100th candidate draw, it is set equal to the maximum).

4.5.3 Levels of the parent/child tree

The implementation of the parent-child relationships and the imposition of exact restrictions are accomplished by assigning every variable a level in the binary tree representing the graph of the parent-child relations. Hence, this information must be provided by the analyst for every variable. The level governs the order in which the sequential regression imputation is done. If the variable does not depend (for any reason) on another variable being modeled first, then it is at the first level, the root of the graph representing the binary tree. Otherwise, a variable must be one level higher than the highest level of any variables on which it depends, so that estimation occurs when the algorithm reaches a node with a binary decision or a leaf of the tree (nodes which are not parents of any variable) where the child variable is not structurally missing. The dependence modeled in the binary tree can be either in the form of a parent-child relationship or constraints. The variable list is then sorted by level (ascending) and missingness (descending) so that all first level variables are imputed or synthesized in a given iteration prior to second level variables, *etc.* In most cases, any variable with either a parent or restrictions of some type will be either a level two variable or higher.

There are a few exceptions. If a parent variable or a variable imposing a restriction is never missing and will not be synthesized, then its child variable or constrained variable can still be at level one for purposes of the estimation.

At the outset of each iteration, the values of all parent variables are stored in a separate file, *orig_parents*. Since a parent variable must be at a lower level in the tree than its associated child variables, in any given iteration, it will be imputed or synthesized before its children. Once a parent variable has been imputed or synthesized, the current iteration file contains the most up-to-date parent values. The previous iteration's values of the parent variable are still in *orig_parents*. However, at this point in the iteration cycle (after a parent has been imputed but before its children have been imputed), the previous iteration's parent values are the ones that correspond to the most up-to-date child variable values. Hence, when the programs reach the point at which they must estimate current iteration models for child variables, they use only observations where the value of the parent variable in *orig_parents* falls in the aforementioned range for the estimation.

At each level and for every variable, fresh model estimation is used to form the posterior predictive distribution. However, when actually imputing values (sampling from the PPD), the programs use the most-up-to-date parent variables to select the observations that will receive values for the children variables. Thus, when the iteration is finished, the parent and children variables all agree again. Child variables only take on values when their parent variables are in the appropriate range and all other observations are set to SAS missing to denote structural missingness.

4.5.4 Grouping and conditioning variables

Finally, as described in sections 4.2 and 4.3, the analyst chooses both grouping variables and conditioning variables. Grouping variables are chosen so that each group meets a minimum size requirement and at the same time contains people who are as similar as possible. In SRMI models, adding additional grouping variables is very costly in terms of computational time so the analyst must seek to make a parsimonious but effective list of variables to use for group stratification. Each unique group, defined by the values of all the variables in the grouping list, has its own posterior predictive distribution. This is the equivalent of fully interacting every grouping variable with every conditioning variable. Conditioning variables are used so that within homogeneous groups, important relationships between the dependent variables and other variables on the file can be preserved.

Problems develop when the grouping variables produce sub-groups that are too small to estimate a statistically reliable PPD. We use the rule that the number of observations in any sub-group must be at least 15 times the number of conditioning variables or 1,000, whichever is greater. To implement this rule, the programs begin with the complete set of grouping variables, form all possible sub-groups, and then check their sample sizes. Sub-groups that are too small are collapsed along specified dimensions and then split into sub-groups again, using a list of grouping variables that is shorter and produces fewer groups. Hence, the analyst actually specifies multiple lists of grouping variables and conditioning variables for each model. Each set of grouping variables is defined by progressively fewer variables as variables are dropped in order to create sub-groups of larger sizes. As variables are dropped from the grouping variables list, they are added to the list of conditioning variables. Hence, each list of conditioning variables becomes

progressively longer. For example, the analyst might originally use *black*, *male*, and *age_cat_expand*, an 11 category age variable, as grouping variables. This would produce 44 groups (2 categories for *black*, 2 categories for *male*, and 11 categories for age). The program would form these 44 sub-groups and check the sample size of each group against the minimum of 15 times the number of conditioning variables. If the analyst included 7 conditioning variables, each sub-group would need at least $\max(1000, 105) = 1000$ observations. If the analyst included 100 conditioning variables, then each sub-group would require at least $\max(1000, 1500) = 1500$ observations. Any sub-group that was large enough would be sent directly to the modeling step using the specified conditioning variables. All groups that were too small would be combined and then split again using a the next set of grouping variables specified by the analyst. In this case the analyst might use only *black* and *male* as grouping variables and then include *age_cat_expand* in the list of conditioning variables that corresponds to this second list of grouping variables. This process continues until all the sub-groups meet the minimum observation requirements or until the list of grouping variables provided by the analyst is exhausted, at which point all groups that are still too small are combined and sent to the regression modeling step.

As with grouping variables, the initial selection of conditioning variables is dependent on the analyst. However each time a set of candidate conditioning variables is included in the model for a particular dependent variable in a particular sub-group, a Bayesian variable selection process is used to reduce the variable list by eliminating variables that are deemed to have weak relationships with the dependent variable, as measured by the Bayes Information Criterion (BIC). The analyst controls the criteria for determining the critical BIC (posterior odds ratio for the model including the variable versus the model excluding the variable) and can make the selection criterion stronger or weaker, depending on the need to keep fewer or more conditioning variables. In version 4.0, we have considerably weakened the critical BIC in order to ensure that important conditioning variables were not dropped from the right-hand side of models.

4.5.5 *Specific variable details*

We have created an Excel workbook with spreadsheets that give the details of the synthetic data creation procedure for every variable on the public use file.¹² The workbook is attached to this report and should be useful to analysts who need information about the methods used for any particular variable in the data completion and synthesis phases. We give the source of each variable (SIPP, IRS/SSA, SSA), whether it contained missing data, whether it was synthesized, what type of model was used to complete missing data, what type of model was used to create synthetic data, and the range of values. We list variables that serve as either parents or children and we specify restrictions, if any, imposed by other variables. We describe any post-processing requirements for the variable, including whether any additional variables need to be created for the final file. Finally, we provide a link to the set of grouping and conditioning variables used in the modeling. We also provide this information in Appendix Table A1 which is included with this report as a separate PDF file. In this section of the report, we describe groups of variables and the modeling techniques used for the group in both the completion and synthesis phases.

Unsynthesized variables Early discussions among committee members produced a list of variables that would not be synthesized: gender, race (black/African-American), three categories of education, marital status, three categories of age, and a link to the record of the spouse at the time of interview. The idea behind unsynthesized variables was that these would enhance the analytic validity of the synthetic file by preserving some basic individual characteristics. Unsynthesized variables, however, also provided a very effective matching strategy for anyone trying to link the new synthetic public use file to the original SIPP public use files. If the unsynthesized variables are used to stratify the sample and if some combinations produce very small groups of people in the Gold Standard file, then an intruder attempting to link synthetic data records to already public SIPP files could match these small groups and might be able to re-identify some individuals in the original SIPP public use files. Thus, this original list of unsynthesized variables was chosen to minimize the number of cells in the Gold Standard file with fewer than 10 people when cross-classified by all the unsynthesized variables.

During this final year of the project, the Census Bureau and SSA conducted lengthy discussions about the possibility of including unsynthesized SSA benefit variables on the file. Although these variables were administrative and hence did not have direct equivalents in the original SIPP survey files, the Census Bureau was concerned that adding more unsynthesized variables to the file would create even more small cells that would allow a user to link across

¹²See `varlists_description_version_4.0.xls` in the appendix to this report.

synthetic implicates. If the synthetic implicates were linked, they could be averaged and something resembling the original record could possibly be re-created. The Census Bureau felt that this possibility presented too much disclosure risk and preferred to keep the number of unsynthesized variables small enough to avoid large numbers of cells with fewer than 10 people.

Discussions between the two agencies produced the following compromise. Gender, marital status, and the spouse-link would remain unsynthesized. In addition, we would add two important SSA benefit variables to the unsynthesized list: type of benefit at time of initial benefit receipt and type of benefit in April 2000. These two categorical variables quantify fact of receipt as well as the reason and are hence the most fundamental of all the SSA benefit variables. Thus, the list of unsynthesized variables in the final version of the synthetic public use files is gender, marital status, initial type of benefit, type of benefit in 2000, spouse initial type of benefit and spouse type of benefit in 2000 (both created using the unsynthesized spouse link), and the spouse identifier variable.¹³ The resulting configuration of unsynthesized variables creates no small cells using only the variables originating from Gold Standard SIPP variables. Furthermore, there are only approximately 130 cells with fewer than 10 individuals when stratifying using the full list, which includes the two SSA-provided type of benefit variables that are not present on any current SIPP public use file. See Table 1 for a full break down of small cells created by various configurations of unsynthesized variables.

The existence of unsynthesized variables required the imposition of some constraints on other variables. In particular, receipt of certain types of benefits imposed constraints on an individual’s age at a given point in time and marital status at the time of the survey imposed constraints on the marital history of an individual. We describe how we handled these restrictions in sections 4.5.5 and 4.5.5, respectively.

Birth date, death date, and dates of benefit receipt One of the most important variables in the file, from the perspective of both disclosure risk and usefulness in analyses, is *birthdate*.¹⁴ It was essential that this variable be adequately protected yet synthesized well enough to reproduce appropriate age distributions for many sub-groups. We used the administrative value of the date of birth (from SSA administrative records) whenever we could. The administrative *birthdate_pcf* was missing in cases where the individual did not have a validated SSN and was completed using the couple-level Bayesian bootstrap described in 4.4. We modeled the variable in the data synthesis phase as a continuous variable with restrictions. If a person received benefits in April 1, 2000 ($tob_{2000} = \{1, 2, 3, 5, 100\}$), we forced the synthetic *birthdate* to be such that the individual would be appropriately old enough for the benefit received. Individuals with retirement benefits had to be at least 62 by April 1, 2000, individuals with aged spouse benefits ($tob_{2000} = 3$) had to be at least 62, individuals with aged widow benefits ($tob_{2000} = 5$) had to be at least 60, and individuals receiving disability benefits ($tob_{2000} = 2$) could not be 65 years old or older. In addition, we restricted draws for the synthetic *birthdate* such that it was no more than a year in either direction from the original administrative birth date (*birthdate_pcf*). So that

$$birthdate_pcf - 365 \leq birthdate \leq birthdate_pcf + 365$$

where we note that date variables are measured in days.

Because *tob_2000* and *tob_initial* are unsynthesized, further consistency restrictions were imposed on *birthdate*. If an individual’s initial benefit types were retired or retired spouse ($tob_initial = \{1, 3\}$) and unsynthesized *date_initial_entitle* is before April 1, 2000, then *birthdate* must be consistent with age at April 1, 2000 greater than 62. The reason for this restriction is that when *date_initial_entitle* is synthesized, there will be support for a synthetic value that is consistent with these types of benefits starting before April 1, 2000. If an individual’s initial benefit types were retired or retired spouse ($tob_initial = \{1, 3\}$) and unsynthesized *date_initial_entitle* is on or after April 1, 2000, then *birthdate* must be consistent with the individual turning 62 (and thus being eligible for these types of benefits) before December

¹³We did make one change with respect to the gender variable that was necessitated by disclosure risk. The Gold Standard contained 5 married couples that had the same gender. Due to the unusual nature of these cases, we could not leave gender and marital status unchanged for these couples without ensuring a link between the synthetic data and the public use SIPP. Hence for these 5 couples, we randomly changed the gender of one of the spouses. We did so in a manner that allowed the weighted counts of males and females in the synthetic data files to remain close to what they were before the gender swaps.

¹⁴In the synthetic data files there is only one birth date variable: *birthdate*. In the Gold Standard file, there are two birth date variables: *birthdate_pcf*, the administrative birth date, and *birthdate_sipp*, the SIPP birth date. The SIPP birth date is only used during the disclosure avoidance analysis.

31, 2002.¹⁵ The same process is repeated for aged widow benefits ($tob_initial = 5$) using an age cut off of 60. Finally, for disabled benefits ($tob_initial = 2$) we reverse this procedure to keep $birthdate$ consistent with being less than 65 years old when this type of benefit is collected. If $tob_initial = 2$ and unsynthesized $date_initial_entitle$ is on or after April 1, 2000, then the minimum synthetic $birthdate$ is May 1, 1935 so that there is support for a synthetic $date_initial_entitle$ on or after April 1, 2000 and the individual would be age-eligible for disability benefits at that time.

The completed data, which are based on the Gold Standard file and which contain either the matching administrative data for the individual and his/her spouse or a complete administrative record (all dates, all earnings, and all benefit variables drawn from the same individual's administrative records and his/her spouse), exhibit some dating inconsistencies that are not due to either the missing data imputation or the synthesis. Because age eligibility restrictions have been imposed in the synthetic data, the synthetic data are cleaner than the completed data; that is, they do not display as many age-related eligibility anomalies as can be seen in the completed data.

The variable $deathdate$ was completed in a similar manner as $birthdate$, using the donor chosen in the couple-level Bayesian bootstrap. This variable was also modeled as a continuous variable during the synthesis phase; however, we also synthesized whether or not the individual died ($flag_deathdate_exist$). The construction of $flag_deathdate_exist$ used the existence of a date of death in the PCF as the indicator of death without modification. The synthetic $flag_deathdate_exist$ is the parent to $deathdate$. $Deathdate$ was restricted such that the earliest possible year of death was 1990.

The following constraints on $deathdate$ obviously only pertain to the cases where the synthetic death indicator is in scope ($flag_deathdate_exist = 1$). In the cases where the completed $flag_deathdate_exist = 1$, we constrained the draw of synthetic $deathdate$ to be within 365 days of the completed $deathdate$. If benefits were received in the month of April 2000 ($tob_2000 > 0$), then the minimum value of the synthesized $deathdate$ is April 1, 2000, since we do not want anyone receiving benefits after death. If there is no benefit amount reported for the entire month of April 2000 ($tob_2000 = \text{SAS missing}$), the initial benefit type is present ($tob_initial > 0$), and the unsynthesized initial entitlement date is before April 1, 2000 ($date_initial_entitle < \text{April 1, 2000}$), then $deathdate$ can be no later than March 31, 2000. Thus, if an individual dies and stops receiving benefits between the initial entitlement date and April 1, 2000, we ensure that the explanation for this loss of benefit is the date of death. If there is no benefit amount reported for the entire month of April 2000 ($tob_2000 = \text{SAS missing}$), the initial benefit type is present ($tob_initial > 0$), and the unsynthesized initial entitlement date is on or after May 1, 2000 ($date_initial_entitle \geq \text{May 1, 2000}$), then the minimum value for $deathdate$ is May 1, 2000. We do this to create support for a draw of the synthetic date of initial entitlement that is consistent with receiving no benefits in the month of April 2000 and the synthetic date of death.

The final date variable that we completed and synthesized was year of initial entitlement to SSA benefits ($date_initial_entitle$). Both completion and synthesis were done in the same manner as the $birthdate$ and $deathdate$ variables. The restrictions on the initial entitlement variable were derived from the draws for the synthetic $birthdate$ and $deathdate$ as well as from the type of benefit variables. If initial type of benefit was retired worker ($tob_initial = 1$), then year of initial entitlement had to be at least 62 years (actually 62×365.25 days) from the synthetic $birthdate$ value. For other types of initial benefits we imposed the following restrictions: at least 62 years old for aged spouses ($tob_initial = 3$), at least 60 years old for aged widows ($tob_initial = 5$), and less than 65 years old for disabled workers ($tob_initial = 2$). Date of initial entitlement had to be before $deathdate$ and before April 1, 2000 if type of benefit 2000 indicated benefit receipt at this point in time. If no benefits were received in April 2000, then date of initial entitlement had to be after April 2000. Hence, date of initial entitlement did not cross the April 2000 boundary. We made two additional restrictions. Because the MBR file did not provide benefit amounts prior to 1962, we did not allow date of initial entitlement to cross the January 1962 boundary. This allowed us to leave the monthly benefit amount variable missing for those with a synthetic (and original) date of initial entitlement prior to January 1962. Finally we restricted draws for $date_initial_entitle$ such that the synthetic value was forced to be no more than 2 years in either direction from the original value.

Administrative earnings After completing missing SER and DER data using the couple-level Bayesian bootstrap, the administrative earnings variables were synthesized in two parts. We first modeled whether the SIPP individual had positive earnings in a given year and then only modeled actual earnings for those with a positive earnings indicator.

¹⁵In Version 4.0 of the Gold Standard and SIPP/SSA/IRS-PUF the SSA MBR data end with calendar year 2002, even though the earnings data end with calendar year 2003. This separation is due to the schedule of extract updates maintained between the Census Bureau and SSA.

Thus, the earnings indicator was the parent variable and the actual earnings variable was the child. We synthesized the earnings indicators using a Bayesian bootstrap, done one year at a time. We used leads and lags for previous and future years as grouping variables as well as demographic variables and summary earnings measures. We began with SER earnings (capped at the FICA maximum) in 1951. Using the bootstrap, we created a synthetic value for every individual for the variable $ser_posearn_{1951}$. For those with $ser_posearn_{1951}=1$, we then used a bootstrap to create a synthetic value for whether each individual had reached the FICA taxable maximum in 1951 ($ser_maxearn_{1951}$). For those with $ser_maxearn_{1951} = 1$, we automatically set $totearn_ser_{1951}$ equal to the maximum. For those with $ser_maxearn_{1951} = 0$, we modeled earnings using our continuous variable techniques, including the two-sided KDE transform. After 1951 was completed, we moved to 1952 and repeated the process. When creating grouping variables for 1952, we used the new synthetic values for 1951 and the completed data for 1953 and after. We moved through the entire array in this manner until the year 1978.

The DER array of earnings begins in 1978. Beginning with this year, we synthesized total earnings. We used a similar process to the one used for the SER earnings except that we synthesized four separate time series: non-deferred total earnings at FICA covered jobs ($nondefer_der_fica_{year}$), deferred total earnings at FICA covered jobs ($defer_der_fica_{year}$), non-deferred total earnings ($nondefer_der_nonfica_{year}$) at non-FICA covered jobs, and deferred total earnings at non-FICA covered jobs ($defer_der_nonfica_{year}$). After each year of DER earnings was synthesized, we calculated SER earnings as the lesser of total non-deferred and deferred earnings at FICA covered jobs or the FICA taxable maximum:

$$totearn_ser_{year} = \min(taxmax, nondefer_der_fica_{year} + defer_der_fica_{year})$$

This process was continued until 2003, the last year of available earnings data.

One final constraint was imposed on the SER and DER earnings arrays. Earnings could only be positive in years where the individual was at least 15 years old and in years up to and including date of death.

Social Security benefits We synthesized two SSA benefit variables: monthly benefit amount for the month of initial entitlement and monthly benefit amount for April 2000. Each of these variables was the child of the corresponding type of benefit variables. Only individuals with a positive initial type of benefit received a synthesized value for the initial MBA ($mba_initial$) and likewise for mba_{2000} . However since neither type of benefit variable was synthesized, the set of people with positive $mba_initial$ and mba_{2000} values was the same in the completed and synthetic data. Both MBA variables were synthesized using continuous variable methods and were restricted such that synthetic values had to be no more than \$50 less than or greater than the original values:

$$mba_initial(completed) - \$50 \leq mba_initial(synthetic) \leq mba_initial(completed) + \$50.$$

and similarly for mba_{2000} .

Once the synthetic data files had been created, we created two additional variables that were direct derivatives of SER earnings: Average Indexed Monthly Earnings ($AIME$) or Average Monthly Wage (AMW) and Primary Insurance Amount (PIA). The $AIME/AMW$ calculation is the method used to summarize a person's lifetime earnings in order to make OASDI benefit calculations. The $AIME/AMW$ is used to calculate the PIA , which in theory tells what benefit a person receives. However, additional rules about spouses, children, family maximums, etc., mean that the actual monthly benefit amount often differs from the PIA . The precise calculations for the $AIME/AMW$ and the PIA depend on a person's gender, date of birth, type of benefit sought, and year of application. The rules governing these calculations are quite complicated (partly because they change a great deal over time) and depend on many things not necessarily observable in our data set. The PIA is an actual variable on the SSA Master Beneficiary File (MBR), but the decision was made by SSA and the Census Bureau not to synthesize this variable or include it on the file, primarily because of concerns that it would be inconsistent with the synthetic SER earnings array. Instead, it was decided that the $AIME/AMW$ and the PIA would be calculated directly from the synthetic earnings using a simplified set of rules.

For individuals who reached age 62 before 1979, we calculated the AMW and for those who reached age 62 after 1979, we calculated the $AIME$. To compute the AMW , we first calculated the number of years between age 21 (or 1951 if later) and age 62, subtracted five years, and multiplied by 12 to get the number of months at risk. We

then summed earnings between age 21 and age 62, dropping the five lowest years. Total summed earnings were then divided by the number of months at risk to give the Average Monthly Wage. There was one exception. For men (but not women) born before 1911, the calculation was performed using the years between age 21 and age 65 because the retirement age for men was three years older prior to 1973. The *AIME* calculation was essentially the same as the *AMW* but earnings were indexed to the year in which the individual turned 60.

Once the *AIME/AMW* had been calculated, the *PIA* was determined by applying the cut-off points and percentages applicable for the year of initial entitlement to benefits. In a given year, $a\%$ of the first X dollars of the *AIME* formed the initial portion of the *PIA*. The $b\%$ of the next Y dollars formed the next portion and $c\%$ of the next Z dollars formed the final portion. The sum of these three portions was the *PIA*. Prior to 1979, the cut-off points stayed constant across years and the percentages changed. Post 1979, the cut-offs changed every year while the percentages stayed constant. We used tables 2.A8, 2.A10, 2.A11, and 2.A16 from the SSA Statistical Supplement 2005 to make these calculations and consulted with Barbara Ling at SSA to clarify details.

It is important to note that we calculated the *AIME/AMW* and the *PIA* for individuals based on the assumption that they were applying for retired worker benefits. We did not make separate calculations for individuals who received disability, spouse, or death benefits. Thus the *AIME/AMW* and *PIA* on the file will not correspond to the *MBA* for types of benefits other than retired worker. However, since the *AIME/AMW* and *PIA* do not contain any additional information and are direct calculations based on other variables in the file, any researcher interested in performing a different calculation may do so. We include these two variables solely for the convenience of retirement researchers.

SIPP time series arrays The synthetic data includes 13 time series of SIPP variables: weeks with pay, weeks part-time, total annual hours, family poverty cut-off, family total income, personal total income, personal total earnings, family welfare participation, family welfare income, private health program participation, private health program income, general health insurance coverage, employer-provided health insurance coverage. In addition, weeks with pay, weeks part-time, annual hours, family income, personal income, and personal earnings have corresponding arrays of indicator variables that serve as parent variables and tell whether the continuous variable takes on a value or not. We use a Bayesian bootstrap to complete and synthesize all the indicator arrays. We then use continuous variable methods to complete and synthesize the remaining variables with the indicators serving as parent variables.

Wealth variables In modeling the wealth variables (total networth, own home indicator, home equity, and non-housing wealth), we create a set of flags to indicate whether the three continuous variables are non-zero. We then use a Bayesian bootstrap to complete and synthesize these three flags together with the home ownership indicator. These four variables are bootstrapped as a group to ensure consistency. We then use the three flags as parents of the three continuous variables. Using our continuous variable techniques, individuals are modeled to have a value of each of the three wealth variables only if the the appropriate flag indicates a non-zero value.

Marital history variables The challenge in synthesizing the marital history variables was to ensure that the historical variables were consistent with the reported marital status and with each other. To accomplish this, we used a Bayesian bootstrap to both complete and synthesize marital history variables. We first bootstrapped a group of variables that summarized the history (*mh_category*, number of marriages, number of divorces, and married at end of history) using marital status as one of the grouping variables. This guaranteed that individuals would receive donated values of *mh_category* and the three other summary variables only from other individuals with the same marital status so no inconsistencies would arise. We then used an additional Bayesian bootstrap for *flag_mar4t* with *mh_category* as one of the grouping variables. Once these variables had been modeled, we created a set of indicator flags that indicated whether the individual should have an age at time of first marriage, duration of first marriage, duration of end of first marriage, duration of second marriage, duration of end of second marriage, duration of third marriage, duration of end of third marriage, and duration of fourth marriage based on the events that occurred in his or her history. Individuals with at least one marriage in their history were modeled to have an age at time of first marriage. Individuals whose first marriage had ended were modeled to have a duration of first marriage and duration of first marriage end. Individuals with at least two marriages were modeled to have a duration of end of first marriage (*i.e.*, time between first and second marriages) and duration of second marriage and so on until the fourth marriage. The age and duration variables were modeled using our continuous variable techniques and were children of the indicator flags.

After the synthesizing was finished, we post-processed these data to create the *mh1-mh7* flags that report the same information as *mh_category*. We used the age at time of first marriage and the duration variables to create the ages at time of each marital history event. To accomplish this, we first summed all the synthetic duration variables to create a total duration and calculated what percentage of the total duration was accounted for by each particular spell. For example, if the individual had 2 marriages, with the second marriage on-going, we calculated what percentage of the total duration was made up of the first marriage duration, time between first and second marriage, and second marriage duration. We took the time period between age at time of first marriage and 2003 (end of our administrative data) and divided it into marital event intervals using the percentages. To continue our example, if age at time of first marriage was 25 and (based on *birthdate*) occurred in 1983, then the total time period was 20 years which would need to be divided between duration of first marriage, interval between first and second marriages, and duration of second marriage. If according to the modeled durations, the first marriage accounted for 50% of the time, the interval between marriages accounted for 25% of the time, and the second marriage accounted for 25% of the time, then age at time of first marriage ending would be 35 (1993) and age at time of second marriage would be 40 (1998).

5 Weight Creation and Synthesis

5.1 Introduction and background

The creation of a unique new public use file that combines SSA/IRS administrative data with extracts from five separate SIPP panels required many special efforts to insure that the final product would be analytically valid. One concern that arose early in the process was how to provide researchers with proper weights for a file that pooled survey respondents from five separate samples. There are design instructions that explain how to combine the official SIPP weights when using panels that contain overlapping years in order to produce estimates that are representative of a known universe at a specific date, but the existing SIPP public use files do not contain the information needed to create a weight that is appropriate for pooling all of the panels into a single analysis. When longitudinal administrative data are linked to these SIPP panels, every observation potentially contributes data to any time period; therefore, the problem of constructing an appropriate weight was integral to permitting these data to be used to make national estimates. In addition, because the different SIPP panels over-sample low income individuals and other targeted demographic groups at different rates, the pooled survey data can only be used to make estimates about the U.S. population if an appropriate weight is used in analyses. Thus, one of the stated objectives of the SIPP/SSA/IRS-PUF project was to create a weight for the five merged SIPP panels where each SIPP person's weight indicated how many persons in the reference population that individual represented. The designated reference population is all individuals age 18 or older in the civilian non-institutionalized U.S. population as of April 1, 2000, the reference date for Census 2000.

In order to determine how many people in the reference population each SIPP person represented, we used the 1996 SIPP sampling plan as our guide and divided the Decennial reference population into the same strata (*i.e.*, groups) from which SIPP individuals were originally sampled. We then located each SIPP individual in the Decennial reference population. Once we knew how many SIPP people were in each stratum, the preliminary weight calculation was straight forward: each SIPP person's weight equals the number of Decennial persons in that particular stratum divided by the number of SIPP persons in the same stratum. For example, if the tenth stratum contained 100 Decennial persons and two SIPP sample individuals, then each SIPP person in the tenth stratum received a preliminary weight of $50=100/2$. The final weight was calculated by raking the preliminary weight to match official U.S. civilian non-institutional population estimates as of April 1, 2000 based on the same control total categories used for the 1996 SIPP weights in the current public use files. The validity of the final weight was tested by computing univariate statistics for key SIPP and SSA variables and comparing them to independently derived estimates from other sources. The results of this testing are reported in Table 2.

In order to locate SIPP individuals in the Decennial reference population, we linked the two data files using the PIK, a unique Census person identifier that replaces the SSN, and which has been added via probabilistic record linking to the Census 2000 micro-data files. For about two-thirds of the individuals in the Gold Standard SIPP file, the PIK link was successful. For the remaining one-third of SIPP individuals, it was not possible to locate an exact match in the Decennial reference population. This occurred either because these SIPP individuals did not provide an SSN to the SIPP survey (and therefore had no PIK) or their PIK did not successfully match to an individual in the Census 2000 micro-data. Of the 263,793 individuals in version 4.0 of the Gold Standard file, 177,165 matched exactly to a Census 2000 reference person by PIK. The other 88,628 SIPP individuals were matched to a Census 2000 reference person using probabilistic record linking.

The strata from the SIPP sampling plan had several levels. The first stratification level (or grouping level) was Primary Sampling Units (PSUs), which were created by grouping geographic counties together. The SIPP Survey Design Branch (SIPPSDB) in the Demographic Statistical Methods Division (DSMD) provided us with a file that assigned geographic counties to PSUs. Large counties were assigned a unique PSU while smaller counties were grouped together to form a single PSU. The second stratification level was by stage-1-clusters, which were simply created by grouping PSUs together. Some PSUs were self-representing, meaning that they were the only PSU in their stage-1-cluster and were sampled with certainty. Other PSUs were non-self-representing, meaning they were grouped with other PSUs and were sampled with probability less than one. The SIPP Survey Design Branch provided us with a file that assigned the 1,928 PSUs to 217 stage-1 clusters. These stage-1 clusters were then used to select PSUs from which individuals would be sampled. Once PSUs were selected, individuals in high poverty strata were over-sampled in each selected PSU. Therefore, our final stratification level was defined by whether an individual was in the high poverty stratum or the low poverty stratum according to the definitions of high and low poverty in the SIPP Sampling Plan. The final stratification which combined the location of an individual in a stage-1-cluster and a poverty stratum

was called a stage-2-cluster. The number of SIPP and Decennial persons in each stage-2 cluster was used to calculate the preliminary weight according to the above formula. Raking was then applied directly to the preliminary weight to create the final weight. Finally, a synthetic version of the weight was created for each of the synthetic implicates.

The rest of this subsection provides the details of this weight creation process. We begin by giving a summary of each of the seven main steps in the process. This summary is meant to give the reader a general idea of how the weight was created before we present the details. Following the summary, parts A-G give careful descriptions of exactly how each step was performed.

5.2 Summary of the weight creation process

Our method for creating an ex-post weight for the merged SIPP panels involved seven steps. Parts A and B describe the method of creating the Census 2000 reference population and dividing it into strata according to the 1996 SIPP Sampling Plan. Parts C and D do the same for the SIPP, describing the method by which the SIPP was divided into strata according to the 1996 SIPP Sampling Plan. Part E describes the method by which each SIPP person was located in the Decennial reference population. Part F describes the creation of the preliminary weight according to the formula mentioned above. Part G describes the creation of the final weight by raking (*i.e.*, adjusting) the preliminary weights to agree with official U.S. population control totals for the sex/age/race/ethnicity demographic breakdown of the reference population, as supplied by the Census Bureau's Population Estimates Division. The next two subsections (5.11 and 5.12) describe some geography and birth date issues that arose during the weight creation process. The next subsection (5.13) discusses the overall evaluation of the Gold Standard weight, and the final two subsections (5.14 and 5.15) describe the creation of the synthetic weight and discuss the results of the analytical validity testing of this weight.

5.2.1 Part A: Creation of poverty stratification variable for Census 2000 records

Part A describes the creation of a poverty stratification variable for Census 2000 records according to original 1996 SIPP stratification rules. Households were assigned to a poverty stratum based on either household income or household composition. For long form households (Sample Census Edited File, SCEF), an income variable was available and households/records were assigned to the high poverty stratum if 1999 household income was below 150 percent of the poverty threshold. For long form respondents for whom income data was not available and for short form respondents (Hundred percent Census Edited File, HCEF), household composition was used to proxy poverty status. A household was assigned to the high poverty stratum if it had any of six characteristics such as a black householder under age 18 or over age 64 (see 5.3 below for the full list of characteristics).

5.2.2 Part B: Creation of stage-2 clusters for Census 2000 records

Part B describes the methods by which counties were assigned to PSUs, PSUs were assigned to stage-1 clusters, and stage-2 clusters were created for the Census 2000 records. This section also describes the manner in which the Decennial reference population was created by only including decennial records that were in the civilian, non-institutionalized U.S. population ages 18 and older on April 1, 2000.

5.2.3 Part C: Creation of poverty stratification variable for SIPP records

Part C is analogous to Part A for the SIPP. It describes the creation of a poverty stratification variable for SIPP records according to the original SIPP stratification rules. Households were assigned to a poverty stratum in the same manner as they were for the Decennial records.

5.2.4 Part D: Creation of stage-2 clusters for SIPP records

Part D is analogous to Part B for the SIPP. It describes the methods by which counties were assigned to PSUs, PSUs were assigned to stage-1 clusters, and stage-2 clusters were created for the SIPP records.

5.2.5 Part E: Matching SIPP individuals to the Census 2000 records

Part E describes the methods by which SIPP persons were located in the Census 2000 reference population. There were 263,793 individuals in the SIPP Gold Standard file, 177,165 of which were matched exactly by PIK to a Decennial record. The remaining 86,628 SIPP records were matched by a probabilistic record linking method to an in-scope Census 2000 record (*i.e.*, a record determined to be in the reference population) in the following manner. Each SIPP

person was first assigned a set of Decennial candidate records (candidates for a match) that agreed exactly with that SIPP record's values for each variable in a set of blocking variables. Then, one of the Decennial candidates was chosen as a match for the SIPP record based on how closely that Decennial record's values agreed with that SIPP record's values for each variable in a set of matching variables. There were two blocking passes through the data. The first blocking pass used 6 blocking variables and 7 matching variables (see 5.8 below for the complete list). Any SIPP record that had 30 or fewer Decennial candidates was considered unmatched and sent through the second blocking pass, which used 3 blocking variables and 10 matching variables.

5.2.6 Part F: Creation of a preliminary weight

Part F describes the calculation of the preliminary weight using Census 2000 stage-2 cluster counts and SIPP stage-2 cluster counts, and the formula above: preliminary weight equals the number of records in Decennial stage-2 cluster divided by the number of records in SIPP stage-2 cluster. This preliminary weight was the same for all SIPP records in a particular stage-2 cluster.

5.2.7 Part G: Creation of final weight

Part G describes the creation of the final weight by raking (*i.e.*, adjusting) the preliminary weights to agree with population control totals for the demographic breakdown of the reference population as provided by the Population Estimates Division. The reference date for the population control totals was April 1, 2000. The list of groups to which the weights were controlled (*e.g.*, black males ages 19-24, black males ages 25-29, *etc.*) was provided by SIPP Survey Design Branch and was the same as the list of population subgroup totals used for raking the original 1996 SIPP weights.

5.3 Part A: Creation of poverty stratification variable for Census 2000 records

We first created the variables that were needed to define poverty status of individuals and households in the SIPP unit frame. The SIPP had four other sampling frames in addition to the unit frame: Area, New Construction, Group Quarters, and Coverage Improvement. However, in the 1996 SIPP panel approximately 80% of records came from the unit frame. Therefore, due to the extraordinary amount of work involved in identifying the stratification rules for the other four sampling frames, we only created the poverty stratification variable according to the unit frame and assumed everyone came from the unit frame. Construction of the necessary poverty-defining variables was different depending on whether the individual completed the Census 2000 short or long form.

5.3.1 Data sources for short-form respondents

For individuals completing the short-form, we took relevant geographic and demographic information from two HCEF data files, namely a person-level file and a block-level file. From the person-level file we obtained indicators for householder, child of householder, spouse of householder, gender, black, Hispanic, age groups (<18, 18-64, >64; and <18, 18-62, >62), birth date, and geography (state, county, approximate tabulation geography). From the block-level file we obtained county, state, population count, housing count, and place code by *geocodecoll* (unique collection block identifier). We then used the person-level data to create a housing-unit file that contained an indicator for family-type housing versus group quarters, a count of persons living in family-type housing, number of children under age 18, householder information (female, black, Hispanic, age: <18, 18-64, >64), and an indicator variable for households with a female householder and no spouse present. Also, in cases where no person was assigned to be the householder (*e.g.*, group quarters have no householder in the Decennial), we assigned the oldest person to be the householder. This file was then merged to each person's record.

5.3.2 Data sources for long-form respondents

Information about long-form individuals came from three SCEF data files: block-level, housing-unit level, and person-level files. From the person-level file we obtained the same demographic and geographic variables as from the short-form: indicators for householder, child of householder, spouse of householder, gender, black, Hispanic, age groups (<18, 18-64, >64; and <18, 18-62, >62), birth date, state, county, and approximate tabulation geography. In addition, we obtained information on education (college, some high school) and income (total annual personal income, 1999). From the block-level file we also obtained the same variables as from the short-form: county, state, population count, housing count, place code by *geocodefull* (unique tabulation block identifier) as well as housing counts and

population counts. Finally, from the housing-unit file we obtained an indicator for family-type housing versus group quarters, count of persons living in family-type housing, number of children under age 18, and an indicator for monthly rent below \$300. We also used the person-level data to create some additional housing-unit information, in particular an indicator for family-type housing versus group quarters, count of persons living in family-type housing, number of children under age 18, householder information (female, black, Hispanic, age: <18, 18-64, >64), and an indicator variable for households with female householder and no spouse present. In cases where no person was assigned to be the householder (*e.g.*, group quarters have no householder in the Decennial), we assigned the oldest person to be the householder. Using the person-level income variable, we created a variable for total annual housing unit income in 1999. All household information was again attached to each person's record.

5.3.3 Data source for MSA variable

The Population Division provided us with a file that included an indicator for "Living in a central city (MSA)". This indicator was merged to the Census 2000 records by state, county, and Census place code. Accordingly, 82,249,968 persons lived in a central city and there were 636 unique central cities.

5.4 Poverty stratum assignment

Households were assigned to strata based on income and household composition. Long-form households for whom an income variable was available were assigned to the high poverty stratum if 1999 household income was below 150 percent of the poverty threshold for that household type. The following list gives the poverty thresholds for various household types.

- if one-person-housing-unit, age of householder ≤ 64 years, and no children under 18 years then poverty threshold 1999(hhpov1999)=8667;
- else if one-person-housing-unit, age of householder > 64 , and no children under 18 years then poverty threshold 1999 =7990;
- else if two-person-housing-unit, age of householder ≤ 64 years, and no children under 18 years then poverty threshold 1999 =11156;
- else if two-person-housing-unit, age of householder > 64 , and no children under 18 years then poverty threshold 1999 =10070;
- else if two-person-housing-unit, age of householder ≤ 64 years, and 1 child under 18 years then poverty threshold 1999 =11483;
- else if two-person-housing-unit, age of householder > 64 , and 1 child under 18 years then poverty threshold 1999 =11440;
- else if three-person-housing-unit and no children under 18 years then poverty threshold 1999 =13032;
- else if three-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =13410;
- else if three-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =13423;
- else if four-person-housing-unit and no children under 18 years then poverty threshold 1999 =17184;
- else if four-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =17465;
- else if four-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =16895;
- else if four-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =16954;
- else if five-person-housing-unit and no children under 18 years then poverty threshold 1999 =20723;
- else if five-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =21024;
- else if five-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =20380;
- else if five-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =19882;
- else if five-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =19578;
- else if six-person-housing-unit and no children under 18 years then poverty threshold 1999 =23835;
- else if six-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =23930;
- else if six-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =23436;
- else if six-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =22964;
- else if six-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =22261;
- else if six-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =21845;
- else if seven-person-housing-unit and no children under 18 years then poverty threshold 1999 =27425;
- else if seven-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =27596;

- else if seven-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =27006;
- else if seven-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =26595;
- else if seven-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =25828;
- else if seven-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =24934;
- else if seven-person-housing-unit and 6 children under 18 years then poverty threshold 1999 =23953;
- else if eight-person-housing-unit and no children under 18 years then poverty threshold 1999 =30673;
- else if eight-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =30944;
- else if eight-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =30387;
- else if eight-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =29899;
- else if eight-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =29206;
- else if eight-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =28327;
- else if eight-person-housing-unit and 6 children under 18 years then poverty threshold 1999 =27412;
- else if eight-person-housing-unit and 7 children under 18 years then poverty threshold 1999 =27180;
- else if >=9-person-housing-unit and no children under 18 years then poverty threshold 1999 =36897;
- else if >=9-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =37076;
- else if >=9-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =36583;
- else if >=9-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =36169;
- else if >=9-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =35489;
- else if >=9-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =34554;
- else if >=9-person-housing-unit and 6 children under 18 years then poverty threshold 1999 =33708;
- else if >=9-person-housing-unit and 7 children under 18 years then poverty threshold 1999 =33499;
- else if >=9-person-housing-unit and >= 8 children under 18 years then poverty threshold 1999 =32208;

When income data were not available for long-form households, household composition was used to proxy poverty status. A household was assigned to the high poverty stratum if it had any of the following characteristics:

- 1) Female householder with children under 18 and no spouse present;
- 2) Living in a central city of a MSA and renter with rent less than \$300;
- 3) Black householder and living in a central city of a MSA;
- 4) Hispanic householder and living in a central city of an MSA;
- 5) Black householder and householder less than age 18 or greater than 64;
- 6) Hispanic householder and householder less than age 18 or greater than 64.

Since short form respondents did not report income, the available household composition was used to proxy poverty status.

- 1) Female householder with children under 18 and no spouse present;
- 2) Black householder and living in a central city of an MSA;
- 3) Hispanic householder and living in a central city of an MSA;
- 4) Black householder and householder less than age 18 or greater than 64;
- 5) Hispanic householder and householder less than age 18 or greater than 64.

There were a total of 285,230,516 Decennial records, 64,493,265 of which were placed into the high poverty stratum, and 220,737,251 into the low poverty stratum.

5.5 Part B: Creation of stage-2 clusters for Census 2000 records

In order to group all Decennial individuals into the same stage-2 clusters for SIPP sampling, we first added SIPP sampling frame information to all Census 2000 records. The SIPP Survey Design Branch provided us with several files and memos containing SIPP sampling information. These files assigned Primary Sampling Units (PSUs) to geographic entities (mostly counties, with smaller counties grouped together to form a PSU); determined which PSUs were in the same risk pool to be sampled (*i.e.*, in the same stage-1 cluster); and reported which PSUs were in actuality sampled. Thus, the SIPP sampling frame information allowed us to begin with state and county information from the Decennial file and assign every Decennial record to a stage-1-cluster. We then combined the stage-1 cluster with the poverty stratum created in Part A and created stage-2 clusters.

5.5.1 *Creation of PSUs*

The original file containing the mapping between state/county and PSUs had 3,141 unique state/county observations and 1,928 unique PSU values. However, at this point we encountered a problem caused by the fact that SIPP sampling for the 1990s panels was based on 1990 geography definitions. Since we were creating weights with a reference point of April 1, 2000 and were linking to Census 2000, we needed to extrapolate the 1990 SIPP sampling frame to the year 2000. We therefore needed to take account of the county changes between 1990 and 2000. During that time period several counties were deleted/added/changed in such a way that their geographic changes needed to be addressed.

- Alaska: Denali (02-068) was created from part of the Yukon-Koyukuk Census Area (02-290) and an unpopulated part of the Southeast Fairbanks Census Area (02-240) in December 1990. Given that there were very few people in the area that was taken from the Southeast Fairbanks Census Area, Denali was assigned the same PSU as Yukon-Koyukuk. Yukon-Koyukuk Census Area and Southeast Fairbanks Census Area had different PSUs, but were in the same stage-1-cluster.

- Alaska: Skagway-Yakutat-Angoon Census Area (02-231) was split to create the Skagway-Hoonah-Angoon Census Area (02-232) and Yakutat City and Borough (02-282) in September 1992. Both new counties were assigned the PSU value of Skagway-Yakutat-Angoon Census Area and its stage-1-cluster code.

- Florida: Dade County (12-025) was renamed as Miami-Dade County (12-086) in November 1997. The county codes just needed to be changed for 2000.

- Montana: Yellowstone National Park (30-113) was annexed to Gallatin (30-031) and Park (30-067) counties in November 1997. Park County and Yellowstone National Park were assigned in 1990 to the same PSU and stage-1-cluster, Gallatin was assigned to a different PSU and stage-1-cluster. Because most people were moving to Park County from Yellowstone National Park and only very few people were living in Yellowstone National Park, no changes were made to the sampling frame, except that the record for Yellowstone National Park was taken out.

- Virginia: South Boston City (51-780) changed to town status and was added to Halifax County (51-083) in June 1995. In 1990 both South Boston City and Halifax County belonged to the same PSU. Therefore the change in county status was irrelevant for the assignment of counties to PSUs. The county code just needed to be changed.

The changes outlined above resulted in 2 additional state/county records and in the deletion of 2 other state/county records.

5.5.2 *Creation of stage-1 clusters*

The SIPP Survey Design Branch provided us with a file that assigned the 1,928 PSUs to 217 stage-1-clusters that were used to select PSUs to be sampled. Memos given to us provided the information about the PSUs that were actually sampled from. We merged that information onto the Census 2000 data by PSU.

5.5.3 *Creation of stage-2 clusters*

The Census 2000 file now held information on the 217 stage-1-clusters and on poverty status. The poverty variable had two values, high and low, and, hence, our final grouping of Decennial records contained 434 different stage-2-clusters.

5.5.4 *Dropping Census 2000 records that were out-of-scope for SIPP samples*

Because of the differing nature of a census and a program survey, we recognized the need to exclude some Decennial records as out-of-scope to be sampled for the SIPP. The SIPP Quality Profile 1998, third edition states:

The survey population for SIPP consists of persons resident in United States households and persons living in group quarters, such as dormitories, rooming houses, religious group dwellings, and family-type housing on military bases. Persons living in military barracks and in institutions, such as prisons and nursing homes, are excluded ... The survey population for the SIPP consists of adults (ages 15 and older) of responding households at the first interview. Each original sample member is followed until the end of the panel or until the person becomes ineligible (by dying, entering an institution, moving to Armed Forces barracks, or moving abroad) or leaves the sample. (page 17)

Several groups of the U.S. population that were counted in the Decennial but were out-of-scope for the SIPP based on the above definition and therefore were not considered when calculating the final weight. Accordingly, the following groups were not counted in the strata for the Decennial files:

1. Residents of the commonwealth of Puerto Rico, and residents of the outlying areas under U.S. sovereignty or jurisdiction (principally American Samoa, Guam, Virgin Islands of the U.S., and the Commonwealth of the Northern Mariana Island). This restriction excluded 3,808,610 persons.
2. Residents living in institutional group quarters: persons residing in correctional and juvenile institutions and nursing homes. This restriction excluded an additional 4,059,039 persons.
3. Residents living in non-institutional group quarters: persons living in military quarters, crews of maritime vessels, and staff residents of military institutions. This restriction excluded an additional 361,815 persons.
4. Children under age 18 (born before April 1, 1982). This restriction excluded an additional 72,145,912 persons.

In total, we excluded 80,375,376 Decennial records because they were out-of-scope for the SIPP samples.

5.5.5 *Census 2000 stage-2 cluster tabulations*

After removing the Census 2000 records that were out-of-scope for the SIPP, we made the appropriate Decennial cell counts for the 434 stage-2-clusters explained above. The 204,885,140 Decennial in-scope observations translated into a 472,016.45 mean cell count. The largest cell contained 4,578,514 observations and the smallest cell contained 8,754 observations.

5.6 Part C: Creation of poverty stratification variable for SIPP records

The creation of the poverty stratification variable for each SIPP record involved similar steps to those undertaken for the Census 2000 records. We first created the necessary variables. When data were available, household income in 1999 was created by summing monthly household income across all twelve months for 1999. The following demographic variables were taken from the earliest wave of the SIPP panel in which they were available for each respondent: birth year, birth month, sex, race, and ethnicity. All other demographic variables used for creating the poverty status of a household or the final weight were taken from the year closest to 2000 for each panel, i.e., the last year of each panel. These variables were: dummy variables for female householder, black householder, and Hispanic householder, age of householder (age categories were <18, 18-64, >64), number of children under 18 in the household, and whether a spouse was present in the household.

We then created the poverty stratification variable for each SIPP record. Individuals were assigned to strata based on either household income or household composition. For Gold-Standard respondents surveyed in the 1996 SIPP panel, 1999 household income was available (81,409 respondents) and they were assigned to the high poverty stratum if 1999 household income was below 150 percent of the poverty threshold for their household type. Thresholds were defined according to criteria used for the Census 2000 records (see 5.3).

For SIPP respondents from the early 1990s SIPP panels or for individuals who were missing from the later waves of the 1996 panel because of attrition, 1999 income data were not available. Household composition was used to proxy poverty status. A household was assigned to the high poverty stratum if it had any of the following characteristics:

- 1) Female householder with children under 18 and no spouse present;
- 2) Black householder and householder less than age 18 or greater than 64;
- 3) Hispanic householder and householder less than age 18 or greater than 64.

It was not possible to assign SIPP respondents to the high poverty stratum based on whether they lived in the central city of an MSA (as was done for the Decennial respondents) because this variable depended upon knowing state, county, and Census place code information for each household, and we did not have Census place code on the internal SIPP file. The final stage of adjusting the weight to correct population control totals within sex, race, ethnicity, and geographic location (see 5.10) handled this problem.

Of the 263,293 total individuals in version 4.0 of the Gold Standard file, 33,868 of them were placed into the high poverty stratum, and 229,925 into the low poverty stratum.

5.7 Part D: Creation of stage-2 clusters for SIPP records

As with Census 2000 records, we used the information provided by the SIPP Survey Design Branch to assign Primary Sampling Units (PSUs) to geographic entities. To assign each SIPP individual to a PSU, we needed state and county information. Unfortunately, county level geography was very difficult to obtain for the 1990-1993 SIPP panels. Given the likelihood that an individual's county had changed between the early 1990s and 2000, we did not invest in obtaining SIPP county information for the early panels. Instead we used the state variable recorded for respondents

during the last year of their panel and then randomly assigned county and the corresponding PSU. For respondents from the 1996 panel, state and county geography was available and PSUs were assigned as they were for Census 2000.

Once SIPP respondents were placed in PSUs, the creation of stage-1 and stage-2 clusters proceeded as outlined in 5.5. At this point, we used the link between the Decennial and the SIPP to flag SIPP individuals who matched to a Decennial record that had previously been determined to be out-of-scope, as explained in 5.5. The Gold Standard version 4.0 file contained 263,793 people, 177,165 of which were matched by PIK (*i.e.*, replacement SSN) to a Decennial record. Of these 177,165 records, 2,229 were matched to a Census 2000 record that was out-of-scope for the SIPP, meaning that these SIPP records received a zero weight in the final weight calculation. The remaining 261,564 SIPP in-scope records were used in calculating the weight. The link between the Decennial and the SIPP essentially served to indicate when a person interviewed in the 1990s had experienced a life-change by 2000 that removed them from the reference population.

After removing the SIPP records that matched to out-of-scope Decennial records, we made the appropriate SIPP cell counts for the 434 stage-2-clusters. The 261,564 SIPP in-scope observations translated into a 602.68 mean cell count. The largest cell contained 4,210 observations and the smallest cell contained 2 observations. The strata with very small numbers of SIPP observations could have presented a confidentiality problem when the weight is used on the SIPP/SSA/IRS public use file. We addressed this issue by synthesizing the weight in the Preliminary PUF 4.0 (see 5.14).

5.8 Part E: Matching SIPP individuals to Census 2000 records

There were 263,793 total SIPP individuals in the Gold Standard file, 177,165 of which were matched by PIK to a Census 2000 record. Of these 177,165 records, 2,229 were matched to a Decennial record that was out-of-scope for the SIPP, meaning that these SIPP records received a zero weight in the final weight calculation. Of these 177,165 records, 4,695 were matched by PIK to more than one Decennial record, because sometimes two Decennial records had the same PIK.

5.8.1 Un-duplication of SIPP-Census 2000 matches

Two match scores were created for each Decennial record. The first match score checked whether the Census 2000 record's date of birth and gender matched exactly to the date of birth and gender for that PIK in the Numident data. The first match score also checked whether the Decennial record's date of birth, gender, and race matched exactly to the same variables in the SIPP record. The first match score went up by 1 anytime the Decennial record matched on a variable (either to the Numident data or to the SIPP record). The second match score checked whether the Decennial record's date of birth, gender, and race were allocated or imputed, and went up by 1 anytime one of these characteristics was not allocated or imputed. After creating these match scores, the Decennial record with the highest first match score was chosen as the correct match for the SIPP record. If the two Decennial records tied on the first match score, the one with the highest second match score was chosen. If they tied on the second match score, one Decennial record was chosen at random as the correct match for the SIPP record.

5.8.2 Matching SIPP to Decennial through probabilistic record linking

The remaining 86,628 SIPP records were matched by probabilistic record linking to an in-scope Decennial record. The first blocking pass used 6 blocking variables and 7 matching variables. The 6 blocking variables were

- a. psu (as defined above)
- b. poverty stratum (as defined above)
- c. male (dummy variable)
- d. black (dummy variable)
- e. Hispanic (dummy variable)
- f. birth year

The 7 matching variables were

- a. birth month
- b. children under 18 (dummy variable)
- c. no spouse present (dummy variable)
- d. female householder (dummy variable)
- e. black householder (dummy variable)

- f. Hispanic householder (dummy variable)
- g. age of householder (<18,18-64,>64)

Each SIPP record was assigned a set of Decennial candidates which agreed with that SIPP record exactly on all six blocking variables. Any SIPP record that had 30 or fewer Decennial candidates was considered unmatched and sent through the second blocking pass, which used 3 blocking variables and 10 matching variables. There were 72,866 SIPP records who had at least 31 Decennial candidates, and these SIPP records were each matched to a Decennial record using the same 7 matching variables as above.

For each matching variable, conditional m and u probabilities were created using the definitions in Fellegi and Sunter (1969):

- m = conditional probability that a SIPP-Decennial match had values for the matching variable that agreed exactly, given that the match was correct;
- u = conditional probability that a SIPP record and a randomly chosen Decennial record within the same set of blocking variables had values for the matching variable that agreed exactly.

The m conditional probabilities were estimated using the 177,165 SIPP records who were already matched to a Decennial record by PIK, and the u conditional probabilities were estimated by randomly assigning to each SIPP record in the first blocking pass one of its Decennial candidates. For both blocking passes (using 6 and 3 blocking variables, respectively), m and u probabilities were first created within cells using 3 blocking variables: psu, poverty stratum, and male. The cells were defined by the complete cross-classification of the three blocking variables psu, poverty stratum, and male. If there was a cell which had at least one SIPP record in the probabilistic record link, but no SIPP records in the set already matched by PIK to Census 2000, then that cell had no m probability. For these cells, m probabilities were estimated using coarser cells, first by only 2 blocking variables: psu and poverty stratum, and finally using no blocking variables. In other words, if a cell created from the complete cross-classification of 2 blocking variables was still missing an m probability because there were no SIPP records in that cell which had already been matched by PIK to a Decennial record, then that cell was assigned an m probability using all the SIPP records that had already been matched by PIK to a Decennial record. Whenever an m probability was created using a coarser set of cells, the u probability was created using the same set of cells. In other words, some m and u probabilities were created within cells that used three blocking variables, some within cells that used only two blocking variables, and some with no blocking variables, but the number of blocking variables used to create the m and u probabilities for a particular SIPP record always agreed.

Once m and u probabilities were created for each matching variable and for each SIPP record, agreement and disagreement weights were created for each Decennial candidate as follows: agreement weight = $\ln(m/u)$ and disagreement weight = $\ln((1 - m) / (1 - u))$. These weights were used to create a matching score for each Census 2000 candidate based on whether the Decennial candidate agreed with the SIPP record on the value of each matching variable. Then, the Decennial candidate with the highest matching score for each SIPP record was chosen as the correct match for that SIPP record.

The matching score was created in the following manner: if a Decennial candidate agreed exactly with its SIPP record on the matching variable, that Decennial candidate's matching score went up by the agreement weight for that matching variable, and if a Decennial candidate disagreed with its SIPP record on the matching variable, that Decennial candidate's matching score went up by the disagreement weight (which was always negative) for that matching variable. A few SIPP records had u -probabilities that were greater than m -probabilities for a particular matching variable (which differed across SIPP records), causing that particular matching variable to have no matching power for that particular SIPP record. In this case, that matching variable was not used in creating the matching score, so the matching score went up by zero whether or not the Decennial candidate agreed with its SIPP record on the matching variable.

Once all Decennial records were assigned a matching score, the Decennial record with the highest matching score for each SIPP record was chosen as the match in the following manner: For each SIPP record that was alone in a cell created from the complete cross-classification of the blocking variables (created from 6 blocking variables in the first blocking pass and 3 in the second blocking pass), and hence had a unique set of Decennial candidates, the Decennial record with the highest matching score was chosen as the match. If two or more records had identical matching scores, one record was chosen at random as the match. For the SIPP records who shared cells (created from the complete

cross-classification of blocking variables) with other SIPP records, it was possible that two SIPP records each had the same Decennial record chosen as the match because it had the highest matching score. When this happened, the Decennial record with the higher matching score was chosen (or chosen at random if the two had identical matching scores), and the SIPP record that had been matched to the Decennial record that was not chosen was sent back through to receive another Decennial record as its chosen match from the pool of Decennial records that had not yet been chosen as a match for any SIPP record. This process was repeated until each SIPP record was matched to a Decennial record, and each Decennial record that had been chosen as a match was unique.

The second blocking pass contained the remaining 13,762 SIPP records who had 30 or fewer Decennial candidates from the first blocking pass, and used 3 blocking variables: *psu90sip*, poverty stratum, and male, and 10 matching variables: black, Hispanic, birth year, birth month, children under 18, spouse present, male householder, black householder, Hispanic householder, and householder's age. *m* and *u* probabilities and matching scores were created as they were in the first blocking pass, and a Decennial match was chosen for each SIPP record in the same manner as well.

5.9 Part F: Creation of preliminary weight

After all SIPP records were matched to Decennial records, a preliminary weight was calculated. In order to calculate this weight, we used the Decennial stage-2 cluster counts from 5.5 and the SIPP stage-2 cluster counts from 5.7. The preliminary weight was calculated using the following formula:

$$\text{prelim_weight} = \frac{\text{number of records in Decennial stage-2 cluster}}{\text{number of records in SIPP stage-2 cluster}}$$

This preliminary weight was the same for all SIPP records in a particular stage-2 cluster. The weights ranged from 163.39 to 23,643.67, with a mean of 783.19.

5.10 Part G: Creation of the final weight

After calculating the preliminary weights we calculated population totals for the newly-weighted SIPP for particular subgroups. Given the discrepancy between these totals and the corresponding totals in the Census 2000, the weights needed to be controlled by population totals. We used a method called iterative proportional fitting to adjust the preliminary weights to reflect correct population totals for certain subgroups. This is the same method used by other Census Bureau surveys to calculate final weights. The list of subgroups used was the same list of population subgroups used to adjust the original 1996 SIPP sampling weights, and was provided to us by Tracy Mattingly from the SIPP Survey Design Branch.

To get the population totals for each subgroup, we used the Population Estimates Base for the U.S. civilian non-institutionalized population ages 18 and older on April 1, 2000 as released on the following Population Estimates web site on June 9, 2005: http://www.census.gov/popest/national/asrh/2004_nat_ni.html. A file containing the population totals used from this web site for April 1, 2000, and a spreadsheets containing the population totals that we calculated for certain subgroups have been supplied as part of this final report.

The iterative proportional fitting the preliminary weights to the population subgroup totals for the following demographic breakdown. We first divided the SIPP into four separate tables by race (black/non-black) and ethnicity (Hispanic/non-Hispanic). Then within each table, the rows of the table were the appropriate ages for that subgroup (provided by Tracy Mattingly) and the columns were male/female. The iterative proportional fitting raked the weighted SIPP tables (weighted by the preliminary weights) to the Population Estimates tables, where the numbers to rake to were both the row and column totals from the Population Estimates tables. The output was a set of adjusted tables. For each age/sex cell in a table, the ratio of the adjusted count for that cell to the unadjusted count for that cell was the factor which was multiplied by the preliminary weight to create the final weight for each individual. The final weights ranged from 30.90 to 32,625.69, with a mean of 780.64.

5.11 Geography issues

5.11.1 Different geography concepts on HCEF and other Census 2000 files

In order to establish the poverty indicator we used information from the HCEF and from other files that were merged onto the HCEF either through person IDs or geography (*e.g.*, central city indicators, PSUs). Merging by person IDs did not pose any problems, but merging by geography did. We were working with an internal HCEF file that had not

yet been converted to the “tabulation” geography concept that the SCEF (as well as the other files) used. Our HCEF file had geography that was on collection geography level, which made it easier for the enumerators to perform the interviews. The SCEF file we used (as well as the files that we received from the pop-division and the SIPP Survey Design Branch) had tabulation geography (which was the geography concept that all the Census 2000 tabulations on the web used, for example). While state information was the same for “collection” and “tabulation” geography, county information could be different. Merging therefore was not straightforward. On our internal version of the HCEF was another geography variable (Current geography). This was not “Tabulation geography” either but matches it reasonably well. We used this variable to merge by geography.

5.11.2 *Changing geography boundaries between the 1990s and 2000*

Geography and, especially, county boundaries changed between 1990 and 2000. There were boundary changes as well as the deletion and creation of new counties during that time frame. This affected the use of SIPP-sampling units, because the SIPP sampling units for the 1991-1996 panels were created using the 1990 Census, and the SIPP sampling units for the 1990 SIPP sample used geography from before the 1990 Census. Also, people moved across county boundaries and therefore were counted in different PSU units in 2000 compared to the time they started participating in the SIPP survey. The changes that were made to accommodate the additions and deletions of counties were written down in detail in 5.5. We have not made any changes for counties that purely changed boundaries.

5.12 **Birth date issue**

Several people claimed to have been born on February 29, 1900. SAS did not accept this date as a leap year. We therefore changed the birth date for these people to February 28, 1900.

5.13 **Overall evaluation of Gold Standard weight**

Our method of creating an ex-post weight for the SIPP-SSA public use file utilizes our link between Census 2000 and the SIPP samples of the 1990s to determine how many people in the U.S. population each SIPP individual should represent. This weight will be a key component of the proposed public use product and will allow researchers to confidently represent the U.S. population as of April 1, 2000.

Table 2, Columns B and C presents the results of testing of the Gold Standard weight. We chose several selected statistics from the 2001 SSA Annual Statistical Supplement and calculated these same statistics using our weighted Gold Standard data. Our weighted Gold Standard file reproduces all of these selected statistics fairly closely. In particular, the number of workers receiving retirement benefits in December of 2000 in the Gold Standard data is lower than the number reported by SSA by only one million. The number of widows and widowers receiving benefits in the Gold Standard is lower than the corresponding SSA statistic by only 300,000, and the number of disabled receiving benefits is higher by only 800,000. The average monthly benefit received by these various sets of workers falls within 3%, 7%, and 6% of the SSA reported average monthly benefit for retired workers, widows and widowers, and disabled workers, respectively. The number of permanently insured individuals in December 2000 in the Gold Standard data falls within 1% of the corresponding number reported by SSA, and the number of wage and salary workers with taxable earnings for 2000 falls within 3% of the SSA reported number. The DER average earnings for 2000 in the Gold Standard is about \$3,000 higher than the DER average earnings reported by SSA, and the SER average earnings for 2000 in the Gold Standard is about \$1,400 higher than the SER average earnings reported by SSA. In general, we believe our Gold Standard weight does a particularly good job of reproducing these selected statistics from the 2001 SSA Annual Statistical Supplement.

5.14 **Synthesizing the weight**

The weights on the sixteen synthetic implicates were quite similar across implicates, allowing many observations to be identified across implicates by the value of their weight. Thus, we decided to create a synthetic weight for each synthetic implicate. We created synthetic weights by taking draws from a Dirichlet distribution to obtain the probabilities of having each possible value of the weight for each person in the data.

The theory for sampling from the Dirichlet distribution is described in Tanner (1996), Gelman et al. (2000) and Minka (2003). Suppose that each observation in the data can take on one of k possible outcomes. Let y be the vector of counts of the number of observations that take on each outcome. The multinomial distribution describes this data

as follows:

$$p(y|n; \theta) \propto \prod_{j=1}^k \theta_j^{y_j},$$

where θ_j is the probability of taking on the j th outcome category; these probabilities sum to one ($\sum_{j=1}^k \theta_j = 1$). The total number of observations is $\sum_{j=1}^k y_j = n$. The conjugate prior distribution for this multinomial distribution is known as the Dirichlet,

$$p(\theta|\alpha) \propto \prod_{j=1}^k \theta_j^{\alpha_j - 1},$$

where the θ_j 's are all nonnegative and again sum to one. The posterior distribution for the θ_j 's is Dirichlet with parameters $\alpha_j + y_j$. We call $a = \sum_{j=1}^k \alpha_j$ the "prior sample size" and we call $n = \sum_{j=1}^k y_j$ the likelihood component, or the "data sample size."

In our application, each person in the data can take on one of 55,552 possible values for the weight.¹⁶ The sum of the weights played the role of the "data sample size, and equaled 204,044,727. We used a noninformative prior distribution by spreading additional observations evenly across the 55,552 cells; this was the "prior sample size."¹⁷ The sum of the "data sample size" and the "prior sample size," is called the "posterior sample size" in the posterior Dirichlet distribution for the cell probabilities.

In practice, we replaced the likelihood counts, y_j , with their expected values. We used the SAS procedure PROC CATMOD to model the expected counts for each of the possible 55,552 cells created by the six strata variables (stage-1 cluster, poverty stratum, male, black, Hispanic, and age category). This procedure performs categorical data modeling of data that can be represented by a contingency table. We supplied the procedure with the weighted cell count data from the completed data, where each observation was a cell in the contingency table created by the complete cross-classification of the six strata variables, and each cell count was the weighted sum of the number of persons in that cell. The procedure used maximum likelihood analysis to estimate a log-linear model and calculate the predicted cell frequencies. We computed the maximum likelihood estimates using an iterative proportional fitting algorithm rather than the usual Newton-Raphson algorithm because it allowed us to obtain the predicted cell frequencies without performing time-consuming parameter estimation. The log-linear model included all six main effects (one for each stratum variable), all two-way interaction effects, and a single three-way interaction effect between the poverty stratum, black, and Hispanic variables.

We took four draws from the Dirichlet distribution for each input contingency table coming from one of the four completed data implicates, giving us a total of sixteen draws, one for each synthetic implicate. Each draw provided us with a vector of 55,552 posterior probabilities (which summed to one) for belonging to each of the 55,552 cells. We then multiplied these probabilities by the "data sample size," 204,044,727, to obtain the final weight value for each cell as a whole, and finally divided by the number of SIPP observations in each cell to obtain the final synthetic weight value for each person in that cell.

5.15 Evaluation of the synthesized weight

Table 2, Columns C and D present the results of comparing the weighted completed data to the weighted synthetic data for the same published SSA statistics as were chosen for the testing of the Gold Standard weight. The results from the synthetic data very closely match those from the completed data. Column E shows that the percentage difference between these statistics for the two types of data is very small, ranging from no difference in the number of disabled workers receiving benefits to 4.4% difference in the number of widows and widowers receiving benefits. More specifically, the estimated number of individuals in the reference population receiving retirement benefits in December of 2000 in the synthetic data is lower than the estimated number in the completed data by 700,000. The estimated number of widows and widowers receiving benefits in the synthetic data is lower than the corresponding statistic in the completed data by only 200,000, and the number of disabled receiving benefits is exactly the same in the synthetic and completed data. The average monthly benefit received by workers in the synthetic data falls within 1% of the average monthly benefit in the completed data for all three types of workers. The number of permanently insured individuals in December 2000 in the synthetic data falls within 2% of the corresponding number in the completed

¹⁶This number of different possible values for the weight comes from the fact that the weight differed only by the values of the following variables: stage-1 cluster, poverty stratum, male, black, hispanic, and age category. There were 217 stage-1 clusters, 2 values for poverty stratum, male, black, and hispanic, and 16 age categories, resulting in $217 * 2 * 2 * 2 * 16 = 55,552$ possible unique values for the weight.

¹⁷By agreement with the Census Bureau Disclosure Review Board, we do not disclose the prior sample size when a Dirichlet prior is used for confidentiality protection.

data, and the number of wage and salary workers with taxable earnings for 2000 falls within 1% of the corresponding statistic in the completed data. The DER average earnings for 2000 in the synthetic data is about \$1,400 higher than the DER average earnings in the completed data, and the SER average earnings for 2000 in the synthetic data is about \$800 higher than the SER average earnings in the completed data. Overall, we have shown that our weighted synthetic data does a very good job of matching our weighted completed data on these selected statistics from the 2001 SSA Annual Statistical Supplement.

6 Analytical Validity

Of primary importance to the success of any synthetic data set is the ability to preserve the univariate distributions of variables and to maintain relationships among variables. In this sense, the modeling done to create synthetic data is different than modeling done in order to predict future outcomes or to analyze cause and effect relationships that are important to policy makers. In creating synthetic data, the analyst's goal is to refrain from imposing prior beliefs about the relationships amongst variables and instead to allow the data themselves to determine the nature of these relationships. Thus, when modeling a particular variable, all other variables can potentially be used as explanatory variables, even when such a relationship might not seem sensible to a social science researcher. In practice, due to feasibility issues, the analyst must choose some subset of variables to go on the right hand side of the predictive regressions but the goal remains to impose as few prior beliefs as possible.

Once the synthetic data are created, however, a different kind of analysis becomes necessary, where prior beliefs become important. Standard economic and demographic models must be tested using the synthetic data and analysts with experience evaluating such results must determine whether the synthetic data are statistically valid. We define statistical validity according to Rubin (1996) as:

First and foremost, for statistical validity for scientific estimands, point estimation must be approximately unbiased for the scientific estimands averaging over the sampling and posited nonresponse mechanisms. ... Second, interval estimation and hypothesis testing must be valid in the sense that nominal levels describe operating characteristics over sampling and posited nonresponse mechanisms. (p. 474)

This definition should be modified to include the phrase “confidentiality protection mechanisms” wherever “nonresponse mechanisms” appears.

Thus in order to assess the quality and usefulness of synthetic data, an analyst must determine what statistics are of interest, calculate these statistics, average them over the implicates of synthetic data, and then compare them to the best estimate of the same statistics from the completed Gold Standard data, which we will euphemistically call the “truth” since it is the best available comparison data. If the estimates are unbiased and the variances of the estimates are such that inferences drawn about the estimates are similar to the inferences in the completed Gold Standard (*i.e.*, “true”) data, then the data are statistically valid.

6.1 Complete data estimation

Interest focuses on a complete data estimand Q which is a function of (X, Y) and has dimensions $(c \times 1)$. This estimand can be any computable, vector-valued function of the data. For example, it could be the average value of Y , many moments of Y , conditional moments of Y , given X , parameters of a model relating columns of (X, Y) , percentiles of the distribution of Y , and so on. The essential feature of Q is that it is computable from complete data on the population and, therefore, is not random. To help clarify the ideas of this section, we will use the example of average income in 1990. If we had complete income data on every individual in the United States, *i.e.*, if we knew every element of Y ($N \times p$) associated with the column representing 1990 income, we could calculate the national average with certainty.

Estimates of Q are random because they are based on D , which involves sampling from the finite population and incomplete observation of Y in the sample. We can only calculate an estimate of the average 1990 income because of the sampling involved with the SIPP and because not all SIPP individuals provided 1990 income data. When all sampled individuals provide data on all p variables, there are no item missing data. However, an estimator of Q is still random because of the sample design embodied in I . Even if all SIPP individuals in our sample reported 1990 income, the sample design of the SIPP would still make the average 1990 income a random variable. We will call the complete data estimator $q(D)$ and its variance estimator $u(D)$. Notice that because of the definition of complete data, q and u depend only on (X, Y_{obs}, I) and not on R . The analyst is assumed to have an inference system for $q(D)$ and $u(D)$. In particular, complete data inference can be based on $(q(D) - Q) \sim N(0, u(D))$, which may be exact or an approximation but is assumed to be appropriate in what follows.

6.2 Inference frameworks using multiple imputation

6.2.1 Missing data only

In the classic Rubin (1987) missing data application, Y_{mis} is imputed m times by sampling from $p(Y_{mis} | D)$, the posterior predictive distribution of Y_{mis} given D . The completed data consist of m sets $D^{(\ell)} = \{D, Y_{mis}^{(\ell)}\}$, where $Y_{mis}^{(\ell)}$ is the ℓ^{th} draw from $p(Y_{mis} | D)$ and is called the ℓ^{th} implicate. Continuing the example of 1990 income, we estimate the posterior predictive distribution of missing 1990 income conditional on everything else we observe about the individual (1991 income, gender, race, marital status, *etc.*). We sampled four times and created four implicates $D^{(1)}$, $D^{(2)}$, $D^{(3)}$, and $D^{(4)}$, each of which consists of original non-missing 1990 income data (D) and imputed 1990 income ($Y_{mis}^{(1)} \dots Y_{mis}^{(4)}$). Inference is based on the following formulae:

$$\text{statistic calculated on each implicate file: } q^{(\ell)} = q(D^{(\ell)}).$$

In our example the function q is the average of 1990 income across all individuals in the sample. This average is calculated separately for each implicate and then averaged across implicates as the next formula indicates:

$$\text{average of the statistic across implicates: } \bar{q}_m = \sum_{\ell=1}^m \frac{q^{(\ell)}}{m}.$$

The statistic \bar{q}_m is the new quantity of interest and will serve as the basis for comparison with the synthetic data. Analytic validity requires that synthetic data reproduce \bar{q}_m , on average, and that inferences made about \bar{q}_m remain the same, as expressed by the confidence interval associated with \bar{q}_m . In order to draw proper inferences, the correct variance measure must be used. The variance of \bar{q}_m has two parts. The first part is commonly referred to as the “between-implicate” variance, defined by the following formula:

$$\text{variance of the statistic across implicates: } b_m = \sum_{\ell=1}^m \frac{(q^{(\ell)} - \bar{q}_m)(q^{(\ell)} - \bar{q}_m)'}{m - 1}$$

The measure b_m tells how much variation has been introduced by the multiple draws from the posterior predictive distribution. The second component of the overall variance of \bar{q}_m is calculated by averaging the within implicate variance across implicates. We define the variance of $q^{(\ell)}$ for each implicate ℓ and the average across implicates as follows:

$$\text{variance of the statistic on each implicate file: } u^{(\ell)} = u(D^{(\ell)})$$

and

$$\text{average variance of the statistic across implicates: } \bar{u}_m = \sum_{\ell=1}^m \frac{u^{(\ell)}}{m}.$$

In our continuing example of 1990 income, $u^{(\ell)}$ is the sampling variance of average income (defined as $\frac{s_{income}^2}{N}$) for each implicate ℓ . The total variance of 1990 income is then calculated as a weighted sum of the between implicate variance and the average within implicate variance, defined as follows:

$$\text{total variance of the average statistic across implicates: } T_m = \bar{u}_m + \left(1 + \frac{1}{m}\right) b_m$$

When n and m are large, inference is based on $(\bar{q}_m - Q) \sim N(0, T_m)$. When m is moderate and the estimator \bar{q}_m is univariate (*i.e.*, $c = 1$), inference is based on $(\bar{q}_m - Q) \sim t_{\nu_m}(0, T_m)$, where the degrees of freedom ν_m are defined as

$$\nu_m = (m - 1) \left(1 + \frac{\bar{u}_m}{\left(1 + \frac{1}{m}\right) b_m}\right)^2$$

Proofs and further details can be found in Rubin (1987, 1996).

6.2.2 Missing and partially synthetic data

In order to analyze synthetic data that were created from data that originally contained some missing values, the missing data imputation and the synthetic data sampling must be done sequentially. First, complete m versions of D by sampling from $p(Y_{mis}|D)$. Denote the m completed data sets as $D^{(\ell)} = \{X, Y_{obs}, Y_{mis}^{(\ell)}, I, R\}$, $\ell = 1, \dots, m$. Let the vector Z ($n \times 1$) denote entities i for which any values of Y_{obs} have been synthesized. So, $Z_i = 1$ if any of the values of $Y_{obs,i}$ have been synthesized. Partition Y_{obs} into Y_{nrep} containing the rows where $Z_i = 0$ and Y_{rep} containing the rows where $Z_i = 1$. Then, for each completed data set, partially synthesize r implicates by sampling from $p(Y_{rep}|D^{(\ell)}, Z)$. Denote the r completed partially synthetic data sets as $D^{(\ell,k)} = \{X, Y_{nrep}^{(\ell)}, Y_{rep}^{(\ell,k)}, I, R, Z\}$, $k = 1, \dots, r$ and where $Y_{nrep}^{(\ell)}$ corresponds to the rows of $(Y_{obs}, Y_{mis}^{(\ell)})$ for which $Z_i = 0$ and $Y_{rep}^{(\ell,k)}$ corresponds to the rows of $(Y_{obs}, Y_{mis}^{(\ell)})$ for which $Z_i = 1$. Note that $Y_{nrep}^{(\ell)}$ contains no synthetic data but may contain missing data imputations whereas $Y_{rep}^{(\ell,k)}$ may contain both missing data implicates (an element of $Y_{rep}^{(\ell,k)}$, say ij , for which item j is missing for entity i but not synthesized; entity i is in this set because $Z_i = 1$ whenever any element of Y_{inc} is synthesized) and synthetic data (an element of $Y_{rep}^{(\ell,k)}$, say ij , for which item j is missing for entity i and is synthesized; entity i is in this set because $Z_i = 1$ and element j element of $Y_{inc,i}$ is synthesized).

As with the case of missing data only, a statistic of interest is calculated for each implicate and averaged across implicates. However, because of the data structure that resulted from first completing missing data and then creating synthetic data, the averaging must account for the different types of implicates. Consider the continuation of the example of average 1990 income. Suppose there are 4 missing data implicates and that 2 synthetic implicates per missing data implicate were generated. In the notation used above, $m = 4$ and $r = 2$, which results in 8 unique data sets. We first calculate average income for each of the 8 implicates:

$$\text{statistic calculated on each implicate file: } q^{(\ell,k)} = q(D^{(\ell,k)}).$$

Then, we average across the 2 synthetic implicates that correspond to a given missing data implicate creating $\bar{q}^{(1)}$, $\bar{q}^{(2)}$, $\bar{q}^{(3)}$, $\bar{q}^{(4)}$ according to the formula:

$$\text{average of the statistic across the synthetic implicates: } \bar{q}^{(\ell)} = \sum_{k=1}^r \frac{q^{(\ell,k)}}{r}$$

Finally, we average across all 8 implicates to create \bar{q}_M . This final average can then be compared to the \bar{q}_m created from the missing data implicates only:

$$\text{average of the statistic across synthetic and missing data implicates: } \bar{q}_M = \sum_{\ell=1}^m \sum_{k=1}^r \frac{q^{(\ell,k)}}{mr} = \sum_{\ell=1}^m \frac{\bar{q}^{(\ell)}}{m}.$$

The variance calculations for data that have been completed and synthesized must also account for the additional source of variation that comes from synthesizing. Thus, we calculate the ‘‘between synthetic implicate’’ variance using the following formula:

$$\text{variance of the statistic due to variation in synthetic implicates: } b^{(\ell)} = \sum_{k=1}^r \frac{(q^{(\ell,k)} - \bar{q}^{(\ell)}) (q^{(\ell,k)} - \bar{q}^{(\ell)})'}{r - 1}.$$

This formula quantifies the variation introduced by differences between two synthetic implicates that were generated from the same missing data implicate, *i.e.*, deviations of the synthetic implicate from the average across both synthetic implicates $q^{(\ell,k)} - \bar{q}^{(\ell)}$. We then average this variance over the missing data implicates:

$$\text{average of } b^{(\ell)} \text{ over missing data implicates: } b_M = \sum_{\ell=1}^m \sum_{k=1}^r \frac{(q^{(\ell,k)} - \bar{q}^{(\ell)}) (q^{(\ell,k)} - \bar{q}^{(\ell)})'}{m(r - 1)} = \sum_{\ell=1}^m \frac{b^{(\ell)}}{m}.$$

The next source of variation comes from the multiple implicates due to missing data completion. This variance is calculated using the deviations of the average for a missing data implicate from the overall average, *i.e.*, $\bar{q}^{(\ell)} - \bar{q}_M$. This is the “between missing data implicate” variance:

$$\text{variance of the statistic due to variation in missing data implicates: } B_M = \sum_{\ell=1}^m \frac{(\bar{q}^{(\ell)} - \bar{q}_M)(\bar{q}^{(\ell)} - \bar{q}_M)'}{m-1}.$$

Finally, the last source of variance comes from the within implicate variance, which is averaged across the synthetic implicates for a given missing data implicate and then averaged across all the implicates according to the formulae:

$$\text{variance of the statistic on each implicate file: } u^{(\ell,k)} = u(D^{(\ell,k)}),$$

$$\text{average variance of the statistic across synthetic implicates: } \bar{u}^{(\ell)} = \sum_{k=1}^r \frac{u^{(\ell,k)}}{r}$$

and

$$\text{average variance of the statistic across synthetic and missing data implicates: } \bar{u}_M = \sum_{\ell=1}^m \sum_{k=1}^r \frac{u^{(\ell,k)}}{mr} = \sum_{\ell=1}^m \frac{\bar{u}^{(\ell)}}{m}$$

The total variance is, once again, a weighted sum of the difference sources of variation—between synthetic implicate, between missing data implicate, and within implicate:

$$\text{total variance of the average statistic across implicates: } T_M = \left(1 + \frac{1}{m}\right) B_M - \frac{b_M}{r} + \bar{u}_M.$$

T_M is the variance used to draw inferences about \bar{q}_M and variation introduced by the synthetic and missing data implicates must not be so large that the inferences will be substantially different from those drawn using \bar{q}_m and T_m . When n , m and r are large, inference is based on $(\bar{q}_M - Q) \sim N(0, T_M)$. When m and r are moderate and the estimator \bar{q}_M is univariate (*i.e.*, $c = 1$), inference is based on $(\bar{q}_M - Q) \sim t_{\nu_M}(0, T_M)$ where the degrees of freedom ν_M are defined as

$$\nu_M = \frac{1}{\left(\frac{\left(\left(1 + \frac{1}{m}\right)B_M\right)^2}{(m-1)T_M^2} + \frac{(b_M/r)^2}{m(r-1)T_M^2}\right)}$$

Proofs and details can be found in Reiter (2004).

6.3 Application to the SIPP/SSA/IRS-PUF

Version 4.0 of the public use file consists of 16 implicates. We created four implicates in the missing data completion phase and then created four synthetic implicates per missing data implicate, thus $m = 4$ and $r = 4$. We chose to focus on two types of statistics—regression coefficients and univariate statistics (means, variances and percentiles) because these are most likely to be of interest to the potential users of our public use file. When showing regression results, we report \bar{q}_m and \bar{q}_M as vectors of regression coefficients. To calculate \bar{q}_m we run the same regression on each of the four missing data implicates and then average the coefficients across implicates. To calculate \bar{q}_M we run the same regression on each of the 16 synthetic implicates and then average the coefficients across these implicates. We also report the variance associated with each average coefficient in the form of vectors that contain the diagonal elements of the covariance matrices T_m and T_M . In the same format we report the standard error (square root of diagonal elements of T_m and T_M), t -ratio (each coefficient divided by the standard error), degrees of freedom (calculated using formulae above), and upper and lower bounds of the 95 percent confidence interval. To show the effect of the two types of implicates on the total variance calculation, we also report the component pieces of the overall variance: diagonal elements of B_M , b_M , \bar{u}_M for the synthetic data, and b_m and \bar{u}_m for the missing data. Univariate statistics are reported in the same manner except the results are scalars instead of vectors.

6.4 Results

6.4.1 General interpretation

When comparing results from completed data to results from synthetic data, there are a number of things to consider. First, and most obvious, is how closely to the point estimates correspond to each other. Regression coefficients and moments of the univariate distribution should be similar between the two data sources. However, this leads to the obvious question: “How similar is similar enough?” To answer this question it is important to compare the confidence intervals surrounding the point estimates. In an ideal situation, the point estimates are very close and the confidence intervals completely overlap, presumably with the synthetic confidence interval being slightly larger because of the increased variation due to synthesizing. Results like this give us confidence that the point estimates really are very similar and that inferences drawn about the coefficients will be the same whether one uses synthetic or completed data. In cases where the point estimates are somewhat further apart, the confidence intervals give us some idea of how far off we are. If there is still some overlap, then the synthetic and completed analyses are not so radically different. In cases where there is no overlap of the confidence intervals, the synthetic variable will need to be carefully examined to determine what might have caused the discrepancy.

Even in cases where the synthetic confidence interval contains the entire completed data confidence interval, we might still be concerned with the relative size of the synthetic interval. If the synthetic point estimate is in the middle of a very large interval, then inferences drawn using synthetic data may be too weak. This could happen because the variables being synthesized cannot be well-modeled and, therefore, each synthetic implicate introduces considerable variation into the analyses that involve those variables. This problem can be improved by the creation of more synthetic implicates. Higher numbers of r implicates would reduce the between r -implicate variance, b_M , and tighten the confidence intervals. It would also solve another potential problem. If b_M is too large in the synthetic data, the overall variance T_M can become negative because the b_M term is subtracted in the total variance formula. A large between r -implicate variance swamps other sources of variation and makes the synthetic total variance undefined. When we have cases like this in our results, we revert to the asymptotic formulae (based on $r = \infty$), and note this in the tables. Essentially we calculate T_M as the weighted sum of the between m -implicate variance and the within variance and do not subtract the between r -implicate variance. Then we treat the coefficients as if they were normally distributed and calculate the confidence intervals using the appropriate critical points from the normal distribution instead of from the t -distribution. In the tables we create an indicator called *flag_dfnotexist* which indicates that we could not calculate degrees of freedom for a t -distribution. In cases where the degrees of freedom are less than or equal to two, we also indicate that degrees of freedom do not exist and use the asymptotic (in r) normal distribution to calculate the confidence interval.

It is important to note one more detail about the univariate and regressions results we present here. We have used the weight that we created by matching individuals in our sample to the Census 2000 micro-data. Hence, in both the completed data and the synthetic data, all the statistics we report are weighted and should be interpreted as representative of individuals from the civilian non-institutional U.S. population age 18 or older as of April 1, 2000.

6.4.2 Summary statistics for OASDI beneficiaries

Tables 3-18 give results comparing means of important earnings and benefits variables by demographic group and type of benefit for individuals who became OASDI beneficiaries during the time period covered by these data (*i.e.*, had date of initial entitlement between 1951 and 2002). Tables 3-10 show results for SER work indicators (positive FICA covered earnings in a year) and SER earnings (total FICA covered earnings up to the maximum). As in version 3.1, the percentage of individuals who worked in a given year is very close, on average, for all the groups and across all the years and the confidence intervals overlap. In addition, average earnings are now much closer for all the groups. For example in 1995 average earnings for white males who retire at some point were \$10,347 in the synthetic data and \$11,012 in the completed data. For white females who retire the correspondence is even closer: \$5,495 versus \$5,566. Particularly strong improvement was made for black males. In 1995, black males who retire at some point earned \$8,856 on average in the synthetic data and \$8,564 in the completed data and there is almost complete overlap in the confidence interval. Synthetic earnings data for this group was particularly problematic in earlier versions so this result represents a significant step forward in our modeling. Charts 1 and 2 show the time trend for labor force participation for the four main demographic groups for individuals who retire at some point. Charts 3 and 4 show the same time trend for earnings for the same groups. Labor force participation and earnings trends are the closest

for white women, followed by black women and black men. White men have a slightly higher discrepancy between synthetic and completed earnings in 1985. Still the trend is the same and other years have closer correspondence.

As shown in Tables 11-12, Total SER earnings summed over all years 1951-2003 and total number of years with positive earnings are also very close for most groups. White females who retire at some point earned on average \$192,468 over this time period according to the synthetic data compared to \$198,303 in the completed data and they worked a total of 26.17 versus 26.69 years. None of the individuals who retire or receive disability benefits have total years off by even a full year when comparing the synthetic and completed data. Total earnings differ by between \$1,000 (black males) and \$25,000 (white males). Table 13 shows that patterns of work are also similar between the synthetic and completed data. This table reports on balances in a “personal account” created by taking 2% of earnings annually from age 21 or 1951 (which ever was later) and compounded it annually at 5% interest until the date of entitlement. If individuals worked in predominantly different decades in the synthetic versus the completed data, we would expect the accumulated totals to be very different. Instead these totals are quite similar, even for white male retirees (less than \$1000 difference).

Tables 14-18 compare the synthetic and complete OASDI benefit variables. Table 14 shows that synthetic year of initial entitlement is very close to the completed value, with the differences being less than 6 months for retirees and disability recipients in every demographic group. The confidence intervals are very tight for this variable and the overlap between synthetic and completed is high. Initial MBA in Table 15 is equally well synthesized. The +/- \$50 restriction helps to ensure that the averages for every category in this table are very close. Retired white males received \$716 in benefits on average in the synthetic data, compared to \$730 on average in the completed data. Table 16 shows similar results for average monthly benefit amount in April 2000. Finally tables 17 and 18 show the average *AIME/AMW* and PIA for the various demographic/beneficiary groups. We remind users of the data that these variables were not taken from the Master Beneficiary File but were calculated according to basic retiree benefit formulae for both the synthetic and complete data. Comparisons of these variables between the two data types is simply another way of summarizing the SER earnings data and judging how well we synthesized the history. On average the *AIME/AMW* is very close: \$725 synthetic versus \$765 completed for white retired females and \$1,693 versus \$1,789 for white retired males. The PIA is \$411 versus \$434 and \$749 versus \$776 for the same groups. These results provide us some level of confidence that our synthetic earnings history will produce valid benefit calculations.

6.4.3 Summary statistics for all workers

Tables 19-28 show comparisons between some of the same synthetic and completed variables described above but for all workers instead of just OASDI beneficiaries. Specifically, Tables 19-26 show percentages of individuals who worked and average earnings for the years 1965, 1975, 1985, and 1995. These comparisons show very close correspondence between the synthetic and completed data. Average earnings for white males in 1995 is \$17,047 using synthetic data and \$17,241 using completed data. Of the white males in our sample, 67.1% had positive FICA covered earnings in 1995 according to the synthetic data versus 67.5% according to the completed data. These results are consistent across years and demographic groups. Charts 5-8 show these trends graphically. For whites, the synthetic and completed time trends lie on top of each other. For blacks, there are a few more differences, in particular earnings for black males seem to diverge a bit in 1995, but generally the time trends are close and show the same pattern.

Tables 27 and 28 compare total earnings and years worked from 1951-2003. The group with the closest correspondence on average between the synthetic and complete data is white females (total earnings of \$211,817 versus \$212,751 and total years 17.76 versus 17.99). The group with the largest difference is black males (\$257,525 versus \$240,933 and 18.99 versus 18.41 years). None of the groups differ by more than half a year in the total number of years worked and both black females and white males differ by less than \$10,000 in total earnings.

6.4.4 Summary statistics by education categories

We next consider means of several important variables stratified by race, gender, and education category. In our analyses of version 3.1, we found that the relationship between education and other variables had not always been well preserved in the synthetic data. In this version of synthetic data, we find some improvements in this respect. Tables 29-34 show means for monthly benefit amount in April 2000, uncapped non-deferred earnings from all FICA-covered jobs in the year 2000, total FICA covered earnings up to the maximum in the year 2000, and percentages of foreign-born, Hispanic, and disabled individuals by race, gender, and education. Both *mba_2000* and *totearn_ser_2000* show

very close correspondence between the synthetic and complete data. For example, calculated using the synthetic data, white male college graduates received, on average, \$786 in monthly benefits in April 2000 compared to \$812 calculated from the completed data. For this same group, the confidence interval also overlaps \$758 to \$814 in the synthetic data versus \$796 to \$828 in the completed data. The same group has FICA covered earnings of \$34,103 on average in the synthetic data compared to \$35,830 in the completed data with complete overlap in the confidence interval. Synthetic and completed average FICA covered earnings are particularly close in the “some college” and “college degree” categories for every demographic group, differing by no more than \$500 in almost every group. Total non-deferred earnings (*nd_der_fica_2000*) have greater differences between the synthetic and completed data where the discrepancies range between \$1,000 and \$5,000 on average. The confidence intervals calculated from the synthetic data for the “high school degree,” “some college,” and “college degree” categories overlap the confidence intervals from the completed data for almost every demographic group. The confidence intervals for the “graduate degree” and “no high school degree” categories are generally very large in the synthetic data and sometimes do not overlap the completed data. These categories contain far fewer individuals and, hence, are more difficult to model; consequently, the synthetic data display more uncertainty.

The percentages of individuals who are foreign-born and Hispanic are also very close for the demographic and education sub-groups. Again the three middle education categories show particularly close correspondence between the synthetic and completed data. For example 9.3% of white males with some college are foreign-born according to the synthetic data compared to 9.4% in the completed data. The synthetic and completed data both give 9.0% of individuals as being Hispanic for the same group. In both cases there is complete overlap in the confidence intervals. In past versions, Hispanic was a particularly difficult variable to synthesize but these results seem to indicate that we have made significant progress modeling this variable. Percentages of individuals who report being disabled in the SIPP are also relatively consistent between the synthetic and completed data. White males are the closest across all education categories (%disabled synthetic - %disabled completed < 1% for all groups except graduate degrees) and black males are the most different (but still %disabled synthetic - %disabled < 2% for all groups except graduate degrees), but in all cases there is significant overlap in the confidence intervals.

6.4.5 Summary statistics by foreign born

We also consider means of the three administrative variables discussed in the previous section, stratified by race, gender, and foreign born in Tables 35-37. Both uncapped and capped earnings from FICA covered jobs are relatively close for non-foreign-born and foreign-born workers in the year 2000. For uncapped earnings (DER), the confidence intervals for white females and white males completely overlap, although the intervals for the foreign-born workers are significantly larger in the synthetic data, as one would expect for the smaller sub-group. For black females the confidence intervals overlap but are shifted up slightly in the synthetic data and the same is true for foreign-born black males, while non-foreign-born black males have a large synthetic confidence interval that completely overlaps. A similar pattern holds for capped earnings (SER). Average monthly benefit amounts in April 2000 show some differences when stratified by foreign-born. The difference in the average benefit between the synthetic and completed data is about \$60 for black female, white male, and black male foreign-born individuals. Even though synthetic *mba_2000* is always within \$50 of the completed value for this variable at the individual level, different individuals end up in a given sub-group because black and foreign-born are synthesized. Hence multiple individuals in the completed data who were not foreign-born may have been synthesized to be foreign-born, and hence may move from one sub-group to another, changing the overall average monthly benefit amount in that sub-group by more than \$50, even though as individuals their personal monthly benefit amounts did not change by more than \$50. For these three groups of foreign-born individuals, there is some overlap in the confidence intervals, although the synthetic data interval is somewhat lower than the completed data interval.

6.4.6 Summary statistics for marital histories

Table 38 shows means and confidence intervals for six marital history variables: number of marriages, percent ever divorced, percent ever widowed, duration of 1st marriage, duration of 2nd marriage, and age at first marriage. The first three variables are nearly indistinguishable on average between the synthetic and completed data, clearly the result of a successful Bayesian bootstrap of *mh_category*. The durations are shorter in the synthetic data than in the completed data for both the point estimates and the confidence intervals by 2-3 years. Age at first marriage is approximately 23 years in both data types. The consistent synthesis of these marital history variables is another major step forward

given that past versions of the synthetic data contained synthetic values that did not even meet minimum consistency standards with the unsynthesized and other synthesized variables.

6.4.7 Age at time of retirement

Of particular interest when considering the synthesis of birth date and year of initial entitlement is whether these two variables are consistent enough with each other to produce an expected distribution of retirement ages. Table 39 gives both weighted and unweighted counts of individuals who retired (*i.e.*, had *tob_initial* = 1) at different ages and Charts 9-10 graph the weighted and unweighted distributions respectively. The first important thing to note is that the completed data have some discrepancies between recorded retirement age and legal retirement age. There are almost 5,000 individuals in our sample whose original administrative birth date and year of initial entitlement imply that they retired between age 61 and age 62. It also appears that in the completed data there are large numbers of individuals retiring at age 62 and at age 64. We had expected the spike at age 62 but thought the later spike would be at age 65. In our synthetic data, we attempt to impose the restriction that retirees must be at least 62 and are successful in all but a few cases. Hence the group retiring between ages 61 and 62 vanishes in our synthetic data. The synthetic data also have a high point at age 62 but then taper off more uniformly across ages 63, 64, and 65. Ideally the counts of individuals retiring at age 63 in the synthetic data might have dropped off more quickly. However the modeling is difficult here because the completed data are not entirely as expected and we are forcing some data consistency that does not exist in the original data. Given our careful modeling of date of initial entitlement and its close correspondence on average between the synthetic and completed data, more research is needed to determine the exact cause of the differences in these distributions.

6.4.8 Selected regression results

We begin our discussion of regression results with Tables 40-43 where the dependent variable is the log of total DER earnings (sum of deferred and non-deferred at FICA and non-FICA jobs) in the year 2000. We ran four separate regressions for each of the major demographic groups: white males, black males, white females, and black females. The closest correspondence between the synthetic and completed regression coefficients is in the education variables which always have the same sign and generally have significant overlap in the confidence intervals. The exceptions for overlapping confidence intervals are usually the graduate degree indicator, not surprising given the results in the means presented earlier. The demographic group with the closest synthetic and completed education coefficients is white males. The coefficient on high school degree only in the synthetic data regression is .214 compared to .230 using the completed data, and for some college, the coefficients are .400 and .431 respectively. In both these cases the confidence intervals in the synthetic data contain the confidence intervals in the completed data. In comparison the high school degree only coefficients for black females are .263 and .347 for synthetic and completed data respectively and for some college the coefficients are .494 and .587. The confidence intervals overlap to a great extent but not completely.

The other SIPP demographic variables, Hispanic, disabled, and foreign-born, are not as consistently similar between the synthetic and completed data but they have improved significantly compared to prior versions of the synthetic data. Foreign-born and disabled always have the same sign and Hispanic has the same sign in the regressions for white males and black females. For white males and females the confidence intervals for foreign-born and disabled overlap, and for black males and females the confidence intervals for all three variables overlap. The magnitudes of the coefficients differ but the confidence intervals give reason to be hopeful that the synthetic data are not producing estimates that are entirely different from the completed data.

The right hand side variables with the most discrepancies in these regressions are the experience coefficients (years of positive SER earnings, with squared, cubed, and quartic terms). While the signs are generally the same and the point estimates of the higher order terms are sometimes similar in magnitude, the confidence intervals do not usually overlap, meaning that the synthetic and completed coefficients are significantly different. Using the synthetic data provides a lower return to experience than using the completed data. For example the coefficient on years of experience for white males is .173 in the synthetic data regression versus .275 in the completed data. For black males the difference is .173 versus .388.

Tables 44-47 show regression results where the dependent variable is log of total SER earnings (capped at FICA maximum) in the year 2000. These results follow the same pattern as the DER earnings. The education coefficients are quite similar between the synthetic and completed data and the confidence intervals overlap to quite a large extent.

The signs for disabled, foreign-born, and Hispanic agree in all four regressions and the confidence intervals for black males and females and white males overlap for Hispanic and foreign-born. In addition they overlap for disabled for black males and females and Hispanic for white females. The experience coefficients are again significantly different in the synthetic and completed data. Not surprisingly, the coefficients on marital status indicators are very similar between the synthetic and completed data. Since marital status was not synthesized and was a grouping variable in the modeling of earnings, this is to be expected.

Table 48 shows results for a regression of the log of the *AIME/AMW* variable on various demographic characteristics. The results for this summary measure of earnings generally show point estimates that are quite close between the completed and synthetic data. The race/gender interaction terms have overlapping confidence intervals except for black females and even in this case the point estimates and the intervals are not very different (-.928 versus -.995). The education coefficients all have overlapping confidence intervals with the exception of graduate degree. The Hispanic and marital status indicators are all very close both in terms of confidence intervals and point estimates. Only disabled shows significant bias. The age coefficients are slightly different between the synthetic and completed data but the confidence intervals do overlap.

Tables 49-52 show results using the log of the monthly benefit amount, first at the date of initial entitlement for retired and then disabled workers, followed by the April 2000 amount for the same two groups of beneficiaries. We believe that the monthly benefit amounts are some of the most accurately synthesized variables and these regression results support this belief. For the initial MBA for retirees, all the confidence intervals overlap with the exception of log of total net worth. Even the disability variable performs well, with the synthetic data coefficient being -.048 and the completed data coefficient being -.039 and very close overlap of the confidence intervals. We also include as a right-hand side variable the percentage of years an individual worked from age 15 to time of retirement/death/end of data, whichever was first. This variable also performs similarly in the synthetic and complete data, causing us to be optimistic about the preservation of relationships between years worked and benefits received. The results are equally encouraging for beneficiaries receiving disability payments. Again almost all the confidence intervals overlap (with the exception of graduate degree) and the point estimates are very close. Particularly reassuring is the fact that age at time of initial entitlement and year of initial entitlement have extremely close coefficients in the synthetic and completed data regressions.

The April 2000 monthly benefit amount regressions in Tables 51-52 continue to show close correspondence between the synthetic and completed data. For retirees, the education coefficients are particularly similar between the two data types with overlapping confidence intervals for all four education indicators. Age in the year 2000 and the race/gender indicators are also essentially statistically the same. Even Hispanic and disabled, which have greater differences in the point estimates, have some overlap in the confidence intervals. Log of total net worth continues to be the largest discrepancy between the two regressions. The results for recipients of disability payments are similar. The point estimates are slightly further apart but there is still significant overlap in the confidence intervals.

Tables 53-54 consider comparisons between regressions results using log of total family income and log of total personal income in the year 1999. These SIPP variables have generally been difficult to model because of the large amounts of missing data due to the combination of five SIPP panels. Income variables for 1999 had to be completed for all individuals in the 1990-1993 panels. Thus the majority of completed data is imputed and not taken from survey responses. This fact contributes to difficulties in modeling these variables for synthesis. For many of the coefficients in these regressions, the synthetic degrees of freedom are missing, either because the total variance was negative or because the degrees of freedom calculated using the appropriate formula was not strictly greater than 2. It appears that the between synthetic implicate variance relative to the between completed implicate variance is too high with respect to these variables. In spite of this fact, the point estimates are often quite close between the synthetic and completed data total family income regressions, especially for marital status indicators, type of benefit receipt indicators, race/gender indicators, year of birth, and the first two education categories. Using the asymptotic formula for the confidence intervals in the synthetic data when the degrees of freedom do not exist, there is overlap with the completed data intervals for almost all variables (college only, graduate, disabled, and total number of kids in the family are the only exceptions). Results for total personal income are similar and even show some improvements over total family income. There is overlap in the confidence intervals for the college only and graduate degree indicators so that now only two sets of coefficients are statistically different - disabled and total number of kids in the family.

Table 55 shows results from a logistic regression of a pension indicator variable on various demographic and

economic control variables. These regression coefficients have similar problems to those in the income regressions, namely the synthetic degrees of freedom often do not exist. Again, in spite of this, the coefficients from the synthetic data regression and the completed data regression generally agree in sign and have overlapping confidence intervals that confirm that the magnitudes are not entirely disparate. Some notable exceptions to this are age and age squared in the year 2000 and the industry and occupation indicators where the confidence intervals have no overlap.

Table 56 shows results from another logistic regression where the dependent variable was an indicator for positive weeks with pay in 1999, taken from the SIPP survey. The results here are somewhat mixed. The race/gender coefficients have very similar point estimates and overlapping confidence intervals. The education coefficients diverge somewhat from the completed and synthetic data regressions. Only some college has an overlapping confidence interval. Total number of kids, foreign-born, and Hispanic all have overlapping confidence intervals and quite close point estimates. An indicator for positive benefit receipt in 2000 does not have an overlapping confidence interval. The coefficient on average log real earnings from the DER, calculated using earnings from 1978-2003, has a different sign and non-overlapping confidence interval between the synthetic and completed data regressions. However the quartile indicators calculated using the distribution of this variable have similar coefficients in the two regressions and overlapping confidence intervals. For percentage of eligible years worked in the SIPP, the point estimate appears to disagree between the synthetic and completed data regressions but the confidence interval does overlap and again the quartile indicators are much closer in magnitude with overlapping confidence intervals. The coefficients on the marital status indicators are consistent between the two data types but coefficients on marital history variables that indicate whether the individual has ever been divorced (*divorced1*) or ever been widowed (*widowed1*) are less consistent, with non-overlapping confidence intervals.

The final four tables of regression results 57-60 use wealth variables from the SIPP as the dependent variables. Table 57 begins with log of total net worth for married couples, with one observation per married couple. Control variables for both members of the couple have been included on the right-hand side of the regression. The results here are not as encouraging as in other regressions. The education indicators for both members of the couple have significantly different effects in the synthetic and completed data regressions. Total number of kids, foreign-born, foreign-born-spouse, Hispanic, and Hispanic-spouse do have overlapping confidence intervals but an indicator for owning a home does not. The indicator for receiving retirement benefits also does not have an overlapping confidence interval although all the other benefit indicators do. Table 58 show results from a regression of log of total net worth for single individuals on characteristics of the individual. Again there is often not overlap between in the confidence intervals. Only the high school degree only indicator has any overlap among the education coefficients.

Tables 59-60 show results for the home equity variables. There is some overlap in the confidence intervals for the education coefficients for the main individual in the couple (chosen as whichever individual had the first sorted person identifier variable, basically random) but the spouse education coefficients are significantly different. There is also overlap for the quartile of average log real DER earnings for both the individual and the spouse. Some of the coefficients on the SSA benefit indicators have overlapping confidence intervals and some do not. Year of birth is very similar in the synthetic and completed data. The results for single individuals are similar.

6.4.9 Univariate distributions of continuous variables

Table 61 examines univariate distributions for all the continuous variables in our sample. The continuous variable synthesis techniques used in this project generally did a very good job of modeling the overall univariate distributions of a variety of variables. The percentiles of the synthesized variables match closely with the percentiles of the corresponding completed variables, capturing the general shape of the distribution; although, very sudden spikes and cliffs in the distributions do get smoothed out a bit. Some of the variables had their synthetic draws restricted to rather narrow windows making the close match not too surprising, but even the variables whose synthesis was unrestricted resulted in very similar univariate distributions.

The three date variables were all restricted to be close to the unsynthesized values (when in scope). Synthetic *birthdate* (restricted to be within one year of administrative *birthdate_pcf*) and synthetic date of initial entitlement (restricted to be within 2 years) are extremely close to their completed counterparts with all the percentiles within a couple months of each other. Synthetic *deathdate* appears to struggle a little bit on the lower end of the distribution, but it turns out that this is do to a quirk in the synthetic weight. Unweighted, the synthetic and completed distributions of *deathdate* are also very similar, but the completed files give zero weight to the people who die before the year 2000. The construction of the synthetic weight did not preserve this characteristic, thus making the weighted percentiles at

the lower end of the synthetic *deathdate* distribution seem significantly lower than the completed *deathdate*.

The MBA variables—MBA in the initial month of benefit receipt (*mba_initial_real*) and MBA in April 2000 (*mba_2000*)—were restricted to be within \$50 of the original amounts, thus it is no surprise that the univariate distributions and means were preserved nearly perfectly.

The continuous marital history variables were synthesized without any constraints. As one can see, this did not affect the quality of the age at first marriage synthesis. The synthetic distribution lies almost exactly on the completed distribution. The duration variables which measure the lengths of all applicable events in the marital history—length of first marriage if ever married (*duration_mar1*), length of single spell after first marriage if the first marriage ended (*duration_end1*), length of second marriage if there was a second marriage (*duration_mar2*), *etc.*—exhibit some of the smoothing that can take place in the synthesis when extremely sharp changes occur in the density of the completed variable. For example, *duration_end1* has an extremely dramatic rise somewhere between the 50th and 75th percentiles in the completed data. The synthetic data matches the 25th and 75th percentiles well, but overestimates the median because it has smoothed this spike out a bit. It is also worth noting that some of these duration variables for second, third, and fourth marriages have very small sample sizes which makes synthesis a little less accurate. Nevertheless, the synthetic and completed distributions for these variables match quite closely except for a little smoothing here and there.

The wealth variables have some of the toughest distributions to synthesize. They are highly skewed and have extreme outliers on the high end of the distribution. For both *homeequity* and *nonhouswealth*, the synthesized variables tend to under-estimate the lower end of the distribution and over-estimate the upper end of the distribution, while *totnetworth* also underestimates the lower end of the distribution but matches the upper end of the distribution. The means, however, look very good, and the general shape of the distribution is preserved.

The DER earnings arrays present some of the same challenges as the wealth variables only to a lesser degree. They also have some very large outliers and are heavily skewed. As a result, the synthetic values display some of the same problems as the synthetic wealth variables, but again, to a lesser degree. The lower ends of the distributions tend to be slightly underestimated and the upper ends slightly overestimated. The deferred earnings arrays have extremely small sample sizes and struggle a lot more than the non-deferred earnings arrays, but once again, the means and general shape of the distributions are preserved very well for all the years.

The SER earnings are capped and, therefore, take away one challenge of extreme outliers and introduce a new problem of a truncated distribution. The cap was modeled by introducing another binomial parent variable indicating whether an individual earned equal to or more than the cap in a given year. If not, the amount was modeled with our continuous variable techniques and the draws were restricted to lie between \$0 and the cap. For the most part, these distributions look very good putting about the same amount of weight at the cap and matching the lower percentiles quite closely.

Finally the continuous SIPP arrays all look quite good at the overall univariate level. Although these variables sometimes exhibited analytical difficulties in multivariate analyses, the general approach used for transforming and modeling continuous variables has done an excellent job of matching the percentiles for almost all of these variables. The weeks worked variables are constrained to lie between 0 and 52, but otherwise the synthesis for all the SIPP arrays was unconstrained.

6.4.10 Counts and percentages of categorical variables

Finally, Table 62 shows weighted and unweighted counts and percentages of some of the basic demographic and benefit variables in the synthetic and completed data. Included variables are: *male*, *black*, *Hispanic*, marital status, *tob_initial*, *tob_2000*, home ownership, foreign-born, education category, age category in 1990, age category at time of initial entitlement, and age category at time of retirement. We include these as a help to those seeking to do basic comparisons between the synthetic and completed data.

7 Assessing Disclosure Risk

7.1 Overview

The link between administrative earnings, benefits data and SIPP data adds a significant amount of information to an already very detailed survey and could pose potential disclosure risks beyond those originally managed as part of the regular SIPP public use file disclosure avoidance process. The creation of synthetic data is meant to prevent a link between these new public use files and the original SIPP public use files, which are already in the public domain. In addition, the synthesis of the earnings data meets the IRS disclosure officer's criteria for properly protecting the federal tax information. Our disclosure avoidance research uses the principle that a potential intruder would first try to re-identify the source record for a given synthetic data observation in the existing SIPP public use files, which were used to create the SIPP component of our Gold Standard file.

In order to test the effectiveness of the synthetic data in controlling disclosure risk, we conducted two distinct matching exercises between the synthetic data and the Gold Standard. Since the Gold Standard contains actual values of the data items as released in the original SIPP public use files, the Gold Standard variables are the equivalent of the best available information for an intruder attempting to re-identify a record in the synthetic data. Successful matches between the Gold Standard and the synthetic data represent potential disclosure risks.

It is important to remember that for an actual re-identification of any of the records that were successfully matched to an existing SIPP public use file, an additional non-trivial step is required. This additional step consists of making another successful link to exogenous data files that contain direct identifiers such as names, addresses, telephone numbers, *etc.* Hence, the results from our matching process are a very conservative estimation of re-identification risk.

The Census Bureau Disclosure Review Board has adopted two standards for disclosure avoidance in partially synthetic data. First, using the best available matching technology, the percentage of true matches relative to the size of the files should not be excessively large. Second, the ratio of true matches to the total number of matches (true and false) should be close to one-half. We have performed two types of matching exercises, probabilistic and distance-based. This section describes the results from both exercises and gives an assessment of the risk of disclosure associated with the synthetic data files.

7.2 Matching based on probabilistic record linking

We begin with the probabilistic record linking experiment. Since the public use files consist of 16 different implicates, one must consider the risk associated with each file. In previous runs of this matching process, similar results were found on the different implicates. The evaluation of disclosure risk described here centers on the risk presented by the publication of one single implicate file (the first synthetic implicate that matches to the first missing data implicate, *i.e.* $m = 1$ and $r = 1$). In view of the results that are described below, we expect that similar results would be obtained for the other implicate files individually. We will, however, need to conduct research to evaluate the disclosure risk presented by the release of all 16 implicate files. In particular, we will evaluate the disclosure risk presented by the file obtained by averaging the variables across all the implicate files. The analysis of the averaged file is currently being conducted.

Probabilistic matching requires caching a set of blocking and matching variables that are common to both files. We implemented one blocking strategy using the unsynthesized variables for blocking. For married individuals we use the unsynthesized variable *male* for each member of the couples. For unmarried individuals we use the two unsynthesized variables, *male* and *maritalstat*. The latter can be either widowed, divorced/separated, or never married ($maritalstat = \{2, 3, 4\}$). In other words, for two records to be a match, they must necessarily have identical values for marital status and gender since these two variables were not synthesized. After this has been determined to be the case, other variables can be compared to determine the probability that two records represent the same person.

The probabilistic record linking was performed using the Census Bureau's internal record linking software, which is maintained by the Statistical Research Division. The discussion in this section describes the technical settings used for that software. We set the blank filter flag equal to 0 so that if the variable is missing, the record will automatically be considered to agree on that field. Matching for the two groups, married and unmarried, was done separately. Blocking variables help to reduce the number of records used for comparison; however, in any given run all records in the same blocking group of the synthetic implicate and the Gold Standard files are compared. Thus, record linking computation

is quadratic with run times dominated by the size of the largest block. In this latest version of the SIPP/SSA/IRS-PUF, the block sizes are very large. For this reason, the matching is done within corresponding segments of the Gold Standard and PUF files. Internally we know when segments of the Gold Standard and PUF files (single implicate) correspond to the same individuals, because we make use of the common artificial person identifier (*personid*) that is on both files. Without the information contained in *personid* (which is not on the actual PUF), an intruder would have to compare many more record pairs to find true matches and would not find any more true matches (the true match is guaranteed to be in the blocks being compared) and would almost certainly find more false matches. For this reason our approach leads to a conservative measure of the disclosure risk.

When the SIPP/SSA/IRS-PUF is finally publicly released there will be no link between the Gold Standard data and the synthetic implicate files. However for testing purposes, we have maintained this link by keeping the common person identifier on the Gold Standard file and the PUF implicate files. Thus, by naming this person identifier in the sequence field of the record linking software, we can check which matched record pairs with a given score are correct matches and which are false matches by comparing this person identifier. When the person identifier is the same, the matching algorithm was successful in finding the person in the Gold Standard file to whom the synthetic data record belonged. When the person identifier is different, the matching algorithm was unsuccessful. This technology is also used for the distance matching discussed in section 7.3.

Automatic searches for matches occur only within those records sharing the same values on the blocking variables. Matches agree exactly on values for the blocking variables and, additionally, they agree on values for the matching variables. An input file to the matching software specifies the agreement criterion for each of the matching variables. Two numbers have to be specified for each of the matching variables. The first number represents the conditional probability that the two records agree on the matching field value given that the two records represent a match, called the m probability. The second number represents the conditional probability that the two records agree on the matching field value given that the two records do not represent a match, called the u probability. This technology was also used in creating the weight; see 5.8.2.

From the agreement criterion, the software computes a score. The agreement score for a match on a particular variable from two comparison records is based upon $\ln(m/u)$. A larger ratio implies a stronger distinguishing power for that matching field. Presumably, the ratio $m/u > 1$. When using Census Bureau matching software for the un-duplication of a file, one is trying to identify specific duplicate pairs, so more precise probability estimates may be helpful. However, when using this software for extracting subsets of plausible matches from a large file, the conditional agreement probabilities can be rough general estimates. To lean towards a more conservative assessment of disclosure risk, we obtained the best possible m and u estimates by using the *personid* variable that is common between the files. Given that the public will not have access to this variable, an intruder trying to match the two files cannot possibly obtain better results using matching software that is at least as efficient as the Census Bureau software.

It is easy to calculate the conditional agreement probabilities $m = \Pr(\text{agreement} \mid \text{match})$ for each matching field, if one knows when true matches occur. This is just the relative frequency of the fields on the Gold Standard and PUF files being equal, call this f_0 . It is also easy to calculate the unconditional probability $\Pr(\text{agreement})$ for each matching field that has a categorical variable. If, for example, X is a categorical variable that can take on 3 possible values, x_1, x_2, x_3 then we obtain the distributions of X in the Gold Standard (GS) and PUF files (implicate 1) and calculate

$$\Pr(\text{agreement}) = \sum_{i=1,2,3} \Pr(X = x_i \mid GS) \Pr(X = x_i \mid PUF).$$

Next it is clear that $\Pr(\text{match}) = \frac{1}{N}$, with N being the common size of both the GS and the PUF files, since for each GS record there is only one PUF record representing the same person. Therefore $\Pr(\text{nonmatch}) = \frac{N-1}{N}$, so given $m = \Pr(\text{agreement} \mid \text{match}) = f_0$, we have

$$\Pr(\text{agreement}) = \frac{f_0}{N} + \frac{\Pr(\text{Agreement} \mid \text{nonmatch})(N-1)}{N}$$

and can solve for $u = \Pr(\text{Agreement} \mid \text{nonmatch})$.

The agreement and disagreement conditional probabilities for those variables used for matching individuals with spouses are shown in Table 63. All matching fields were assigned the exact matching comparison type. This caused

the program to assign full agreement/disagreement scores according to whether the fields agree or disagree. The corresponding agreement probabilities for single individuals are just slightly different and are shown in Table 64.

These probabilities are used to calculate the scores given to this variable when it agrees or disagrees. The agreement score is defined as $\ln(\frac{m}{u})$. The disagreement score is defined as $\ln(\frac{1-m}{1-u})$. For example, the full agreement score for a “c-match” on Hispanic is $\ln(\frac{0.888222038}{0.817697432}) \approx 0.08$. The disagreement score is $\ln(\frac{1-0.888222038}{1-0.817697432}) \approx -.50$.

The software compares each matching field, decides whether the field agrees or not, and then assigns the appropriate score to the field based on the user supplied m and u probabilities. Next, a cumulative match score is calculated by summing the scores across all the matching variables. This cumulative score is used to decide whether two records match. It is compared to the cutoff values provided by the user and if it passes the stated threshold, a match is declared. The influence of a one variable relative to another on this cumulative score is controlled by the relative matching and non-matching agreement probabilities specified by the user, but in this case based on actual calculations from the relevant files. The non-matching agreement probability essentially tells how often a field will agree at random across two files. A high value for this probability will reduce the importance of this variable in the matching by causing the agreement score to be lower. This is desirable because if the field is likely to agree at random, any match in values between two files is less likely to signify a true match. At the same time, a high non-matching agreement probability causes the disagreement score to be less negative or smaller, meaning that the penalty for not matching on this variable is not as high. In contrast, the relative matching agreement probability tells the importance of this variable compared to other variables in determining whether two records are a match. A high matching agreement probability means that a match on this field is crucial to determining an overall match between two records. Thus a high value for m produces a high agreement score. It also produces a more negative or higher disagreement score, more severely penalizing non-matching in this field. Consider the example of the variable *flag_mar4t*, which is used to identify individuals who reported more than three marriages. When two records agree on this variable, and they are a match, the cumulative matching score increases by 5.317686217. If the records are not a match, but agree on this variable, then the cumulative score decreases by -4.609063992 .

The output cutoff flag for the cumulative matching score provides the comparison points for the matching score. In our testing we declare any pair of records with a cumulative score between -20 and 20 to be a potential match. That is, we consider matching two records whenever their agreement score exceeds -20 even though most applications of probabilistic record linking use a positive cut-off for the automatic selection of potential matches. Thus, we declare records to be candidate matches based on an aggressive matching strategy. From either Table 63 or 64 we can see that the total matching scores cannot be outside of this range. Essentially, we allow every record in the synthetic file to have candidate matches in the Gold Standard. The output files are sorted by decreasing cumulative agreement score; then, the best two matches are kept. Finally, the proportions of true matches and the ratio of true to false matches are obtained.

The number and proportions of false and true matches, for each of the segments of the file, are given in Tables 65 and 66. The number of true and false matches in each segment are reported in column 3 and sum to equal the total number of records in each segment. The ratio of true to total matches and false to total matches gives the percentage of true matches and false matches in each segment and is reported in column 4. In Table 65, there are no data segments that have a true match rate over 1% and the ratio of true to false matches is extremely low. In Table 66, the percentages of true matches are slightly higher but the highest value is still just over 1% (1.18% for segment 2).

7.3 Distance matching

Distance-based record linking is another common approach to estimating the risk of disclosure in micro data. In recent work, Domingo-Ferrer, Abowd and Torra (2006) use distance-based methods to re-identify records on two synthetic micro-data samples. They find that distance-based metrics perform similarly to (if not better than) the more commonly used probabilistic methods. Their work suggests that re-identification exercises should also include distance based methods. The broader the selection of methods used, the more informed the analyst is of the risk of disclosure. In particular, it is important to understand which methods pose the largest threat. Domingo-Ferrer et al. (2006) conduct similar comparisons of distance-based and probabilistic record linking methods.

Our tests consider the case of an intruder who uses distance-based re-identification to match the source records from the Gold Standard to synthetic SIPP/SSA/IRS-PUF observations. Such re-identification methods calculate the distance between a given record in the Gold Standard and every record in the synthetic implicate. The j closest records

are then declared potential candidates for a match to the source record. In our analysis we consider $j = 3$.

Our distance-based re-identification proceeds in two stages. First we split both the Gold Standard and the first synthetic implicate ($m = 1$ and $r = 1$) into groups based on the unsynthesized variables. In this case, marital status and *male* are the only two unsynthesized variables. We next split each blocking group into smaller segments of approximately 10,000 observations in order to decrease the processing time, which is quadratic in the size of the largest files compared. We performed the segment split on both the Gold Standard and synthetic files so that the correct match in the Gold Standard was always in the same block and segment of the synthetic data used for comparison. In other words, we forced the segmentation of the files to guarantee that the correct match could always be found in the block/segments being compared. This is the same assumption as we used in section 7.2 to segment the comparison files in that analysis. The segmentation of the blocks uses our prior knowledge of which records are actual matches and hence our matching results are conservative—overestimates as compared to a distance record link that could not segment the comparison files because the intruder did not have access to the true *personid*. After splitting the data into blocking groups and segments, we then calculate the distance between a given Gold Standard record and every record in the synthetic file in its corresponding blocking group and segment using a set of 163 matching variables. The three closest records are then declared possible matches.

We use four distance metrics. Each metric is a special case of either Mahalanobis or Euclidian distance. Before formally defining the distance, we first define some notation. Let A and B represent the two data sets being matched. For our purposes, conceptualize the block and segment of the Gold Standard as the A file and the block and segment of the synthetic implicate as the B file. Denote α as the vector of 163 matching variables from an observation in the A file and β as the analogue for the B file. Given this notation we define the distance between a given vector α in the A file and a given vector β in the B file as follows:

$$d(\alpha, \beta) = (\alpha - \beta)^t [Var(A) + Var(B) - 2Cov(A, B)]^{-1} (\alpha - \beta)$$

We consider four specific cases of the general distance. In the first case we assume that the intruder can properly calculate the $Cov(A, B)$. We denote this distance *MAHA1*, and note that it is a true Mahalanobis distance; hence we expect that this distance measure will give us the highest match rates since it uses all of the available information, including the correct covariance structure of the errors in synthesizing all 163 variables. In the second case, we assume that the $Cov(A, B) = 0$. This is equivalent to assuming that we do not know how to link the observations across the A and B files and cannot compute $Cov(A, B)$. A real intruder would not have access to $Cov(A, B)$. We denote the second distance *MAHA2*, and note that it is a “feasible” Mahalanobis distance. In the third case, we assume $[Var(A) + Var(B) - 2Cov(A, B)] = I$, where I is the identity matrix. We denote the third measure as *EUCL1*, which is a Euclidian distance with unstandardized inputs. For the fourth measure, we transform all of the matching variables in the A and B files to $N(0, 1)$ variables. Call the transformed files \tilde{A} and \tilde{B} . We then calculate the distance using $[Var(\tilde{A}) + Var(\tilde{B}) - 2Cov(\tilde{A}, \tilde{B})] = I$. We denote this fourth metric *EUCL2*, and note that it is a standardized Euclidian distance.

Tables 67-68 shows the results of the re-identification exercises for each of the four metrics. Table 67 shows the results using the Mahalanobis distance measures and Table 68 shows the results for the Euclidian distance measures. For each metric there are six columns. Match rate 1 (closest two records in A and B), match rate 2 (second closest two records in A and B), ratio of 2/1, match rate 3 (third closest two records in A and B), ratio of 3/2, and ratio (3+2)/1. Match rate j is calculated as the number of successful matches within a blocking group based on the j th closest observation divided by the total number of observations in that group (multiplied by 100 to convert to percentages). For example, match rate 2 is calculated as the number of successful matches within a blocking group and segment based on the second closest observation divided by the total number of observations in that group (multiplied by 100 to convert to percentages).

We first note that match rate 1 finds the highest rate of re-identifications. This implies that choosing the closest record using the indicated distance metric is more likely to find true match than choosing the second or third closest record. We further note that the highest match rate among all blocking groups is only 2.91%. Thus, an intruder who defined the closest- distance record as a match would correctly link 1.09% of records overall in the synthetic files and less than 3% in the worst-case sub-group.

The three ratio columns give us a sense of how much better the closest match does than the second and third best matches. Ideally, we want to ensure that if an intruder looked at the top three matches, he or she would face sufficient

uncertainty about which one was the correct match. If the second closest record is exactly as likely to be the correct match as the closest record, then the ratio of match rate 2 to match rate 1 would be unity. If this ratio is less than one, then the closest record is more likely to be the correct match. If this ratio is greater than one, then the second closest record is more likely to be the correct match. The other ratio columns have the same interpretation. For the *MAHA1* metric, the column Ratio (3+2)/1 ranges from 0.79 to 1.12. This suggests that the 2nd or 3rd closest matches are almost as likely to be correct as the closest match. The totals in the last row are essentially weighted averages of each column where the weights are the percentage of records in each group.

As expected, the *MAHA1* metric produces the highest match rates. The highest match rate for the *MAHA2* metric, perhaps the most likely to be used by an intruder, is 2.2% and the ratio of (3+2)/1 is very close to unity for every sub-group. The Euclidian metrics are very similar to the *MAHA2* metrics with the overall match rate not exceeding 1.2%, the highest sub-group match rate less than 2.4%, and the ratio of (3+2)/1 generally being very close to or slightly higher than unity.

8 Using Synthetic Data

Many potential users may be concerned about how to begin using synthetic data and multiple implicate files. In this section we give some suggestions and advice for using these data sets to perform analyses and apply the formulae described in 6.

We suggest that users begin with one synthetic implicate and write code to prepare variables and verify the specification of statistical models for this single data set. Since all the synthetic implicates are identical in terms of file structure, number of records, variables names, *etc.*, any code that works on one implicate also works on the remaining implicates. Users can debug their models and, once they are satisfied with the programming specification, run the model on all 16 implicates. In this sense, synthetic data are no different from any other micro-data set. Analyses are run in exactly the same manner but are repeated multiple times. We recommend saving analysis results such as regression coefficients or summary statistics in a data set that can be manipulated on its own. This will be useful for combining results. We also recommend that users base all their statistical inferences on the proper combining formulae. That is, we do not recommend that users conduct statistical specification searches on a single implicate and then estimate “final” standard errors with the proper formulae. The statistical inference theory that underlies partially synthetic data with multiple imputation relies on the multiple analyses, conducted on independently drawn implicates, to reflect the model uncertainty inherent in the original confidential data.

Each synthetic implicate has two variables that control the relationship between implicate files. The variable *m_implicate* tells which completed data implicate served as the starting basis for this particular synthetic data implicate. The variable *r_implicate* gives the synthetic implicate number for the file. There are four completed data implicates, so the variable *m_implicate* ranges from 1 to 4. There are four synthetic implicates per completed data implicate so the variable *r_implicate* ranges from 1 to 4 also. The first synthetic implicate will have *m_implicate* = 1 and *r_implicate* = 1, the second synthetic implicate will have *m_implicate* = 1 and *r_implicate* = 2, and so on, until the fifth synthetic implicate, which will have *m_implicate* = 2 and *r_implicate* = 1. In this manner a user can tell which synthetic implicates stemmed from the same completed data implicate. This information is necessary in order to apply the combining formulae.

Any statistic of interest to a researcher can be calculated from the synthetic data by calculating it once per synthetic implicate and then averaging across the 16 implicates. If the researcher wants to know average earnings in a given year, he or she should calculate the average in each of the 16 implicates using standard methods and then calculate the simple average these 16 separate means to get one grand mean. If the researcher wants to know the variance of earnings in a given year, he or she should follow the same procedure: calculate the variance in each implicate and then calculate the simple average these 16 statistics to get one grand variance. Note, and this is very important, the grand mean of the variances is just one component of the estimated total variance required to compute a confidence interval for average earnings. The complete formula is contained in section 6. Point estimates for any statistic of interest from regression results to moments or percentiles of a distribution can be obtained in this manner. In the standard combining formulae, every implicate is equally weighted, so simple averaging is all that is required.

The calculation of the estimated total variance of a statistic of interest, from which one might compute a confidence interval or test statistic, is more complicated but still can be performed with standard software. In addition to the statistic of interest, the user should save the estimated sampling variance of this statistic for each of the 16 implicates. For example, if calculating one mean per implicate, the user should calculate the sampling variance of the mean once per implicate.¹⁸ The within-implicate sampling variances are then averaged to estimate the average within-implicate variance, one component of the total variance. The user must then make use of the *m_implicate* and *r_implicate* variables to calculate the between-completed-data-implicate variance and the between-synthetic-data-implicate variance according to the formula in 6.2.2. The user first calculates the variance of the statistic across the four *r* implicates associated with a particular *m* implicate. There will be four of these variances: one per completed data implicate. These four variances are then averaged to give the overall between-synthetic-data-implicate variance. The user then calculates the mean of the statistic of interest for all the synthetic implicates associated with a particular completed data implicate. Again, there will be four of these means. The between *m* implicate variance is then calculated as the average of the squared deviations of these four means from the overall grand mean. If the statistics

¹⁸The reader is cautioned to be certain to perform all calculations on variances and not standard deviations. To compute a standard deviation or standard error, the square root operation should be performed on the total variance that has been computed by combining all of the component variances appropriately.

of interest are saved in a data set, these calculations can be easily performed. The variance pieces are then combined to create the total variance and calculate degrees of freedom. In the case that the total variance becomes negative, we recommend not subtracting the between-synthetic-data-implicate variance when calculating the total variance. The confidence interval can be calculated using the asymptotic assumption of normality instead of the finite sample t -distribution.

When presenting research results, users should not report the result from a single synthetic implicate. This is not an accurate representation of either the point estimates or their associated variances. This is especially important when comparing synthetic and completed data in order to determine analytic validity. No synthetic implicate can be judged for accuracy as a stand-alone file. It must be considered in conjunction with the other synthetic data sets. Likewise, all implicates of the completed data must be used together in the manner described above in order to create a comparison basis.

9 Conclusion

Given the length and scope of this project, it is perhaps beneficial at this point to consider what has been accomplished. This collaboration between four government agencies has produced several new data products and advanced the body of knowledge on missing data imputation, assessing the validity of automated statistical modeling, disclosure avoidance techniques, and disclosure risk analysis. In the past six years, we have produced a highly useful compilation of SIPP data that combines five separate panels with edited administrative data from IRS and SSA, a weight to allow meaningful analysis of these combined panels, a set of files that multiply impute all missing data, and a set of synthetic data files that meet disclosure standards of the Census Bureau, the Internal Revenue Service, and the Social Security Administration. For the first time in 30 years, it appears that it will be possible to release lifetime earnings histories taken from administrative records, an accomplishment that will be of enormous benefit to the research community and the general population. This project has been a model for what inter-agency cooperation can accomplish by pooling the expertise of researchers from the Census Bureau, IRS, SSA, and CBO.

When we began this project, there was a great deal of uncertainty over whether synthesizing techniques could produce micro-data that would preserve relationships among variables and mitigate disclosure risk. In fact, almost none of the enhanced theory or experience with these methods required to complete the project existed. Based on the results at this point, we feel that both these questions can be answered in the affirmative. It is now imperative that outside users be given a chance to test these synthetic data and that the agencies involved develop a system for validating outside results using the Gold Standard in order to promote general confidence in the methods and to permit quality improvements. This process will help us to discover remaining flaws in the synthetic data and improve the synthesizing process, both of which will enable the collaborators to provide useful future updates to this data product, as funding resources permit.

Bibliography

- Abowd, John M. and Simon D. Woodcock. 2001. Disclosure Limitation in Longitudinal Linked Data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, ed. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North Holland pp. 215–277.
- Albert, A. and J.A. Anderson. 1984. “On The Existence of Maximum Likelihood Estimates in Logistic Regression Models.” *Biometrika* 71:1–10.
- Domingo-Ferrer, Josep, John M. Abowd and Vicenc Torra. 2006. Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment. In *Privacy in Statistical Databases*, ed. J. Domingo-Ferrer and L. Franconi. Springer-Verlag p. forthcoming.
- Domingo-Ferrer, Josep, Vicenc Torra, J.M. Mateo-Sanz and F Sebe. 2006. Empirical disclosure risk assessment of the IPSO synthetic data generators. In *Monographs in Official Statistics-Work Session on Statistical Data Confidentiality*. Eurostat.
- Fellegi, Ivan P. and Alan B. Sunter. 1969. “A Theory for Record Linkage.” *Journal of the American Statistical Association* 64:1183–1210.
- Gelman, A. B., J. S. Carlin, H. S. Stern and D. B. Rubin. 2000. *Bayesian Data Analysis*. Chapman and Hall.
- Little, Roderick J.A. 1993. “Statistical Analysis of Masked Data.” *Journal of Official Statistics* 9(2):407–426.
- Minka, T. P. 2003. Bayesian inference, entropy, and the multinomial distribution. Technical report Microsoft, Inc. available at <http://research.microsoft.com/minka/papers/minka-multinomial.pdf> (October 30, 2006).
- Raghunathan, T.E., J.P. Reiter and D.B. Rubin. 2003. “Multiple Imputation for Statistical Disclosure Limitation.” *Journal of Official Statistics* 19(1):1–16.
- Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk and Peter Solenberger. 1998. “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models.” Survey Research Center, University of Michigan.
- Reiter, Jerome P. 2003. “Inference for Partially Synthetic, Public Use Microdata Sets.” *Survey Methodology* 29:181–188.
- Reiter, Jerry P. 2004. “Simultaneous use of multiple imputation for missing data and disclosure limitation.” *Survey Methodology* 30:235–242.
- Rubin, Donald B. 1981. “The Bayesian Bootstrap.” *The Annals of Statistics* 9:130–134.
- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Rubin, Donald B. 1993. “Discussion of Statistical Disclosure Limitation.” *Journal of Official Statistics* 9(2):461–468.
- Rubin, Donald B. 1996. “Multiple Imputation after 18+ Years.” *Journal of the American Statistical Association* 91(434):473–489.
- Tanner, Martin A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Third ed. Springer-Verlag.
- Woodcock, Simon D. and Gary Benedetto. 2006. Distribution-preserving statistical disclosure limitation. LEHD technical paper tp-2006-04 U.S. Census Bureau. <http://lehd.dsd.census.gov/led/library/techpapers/tp-2006-04.pdf> (October 31, 2006).

A Appendix

The appendices to this report have been included as separate files. The list is provided here.

- Technical Description of the Creation of the SIPP/SSA/IRS Gold Standard Files and the SIPP-SSA-IRS PUF Version 4.0 (PDF file)
- Appendix Table A1 Detailed Variable Information (PDF file)
- Varlists_description_version_4.0.xls (Excel workbook)
- CompletionSynthesisTables1.xls (table 1) - Completion and Synthesis (Excel workbook)
- WeightsTables1.xls (table 2) - Weights (Excel workbook)
- AnalyticValidityTables1.xls (tables 3-18) - Analytical Validity (Excel workbook)
- AnalyticValidityTables2.xls (tables 19-28) - Analytical Validity (Excel workbook)
- AnalyticValidityTables3.xls (charts 1-8) - Analytical Validity (Excel workbook)
- AnalyticValidityTables4.xls (tables 29-34) - Analytical Validity (Excel workbook)
- AnalyticValidityTables5.xls (tables 35-37) - Analytical Validity (Excel workbook)
- AnalyticValidityTables6.xls (table 38) - Analytical Validity (Excel workbook)
- AnalyticValidityTables7.xls (table 39, charts 9-10) - Analytical Validity (Excel workbook)
- AnalyticValidityTables8.xls (tables 40-60) - Analytical Validity (Excel workbook)
- AnalyticValidityTables9.xls (table 61) - Analytical Validity (Excel workbook)
- AnalyticValidityTables10.xls (table 62) - Analytical Validity (Excel workbook)
- DisclosureTestingTables1.xls (tables 63-66) - Disclosure Testing (Excel workbook)
- DisclosureTestingTables2.xls (table 67-68) - Disclosure Testing (Excel workbook)

Table 1: Small Cell Consequences of Selected Combinations of Non-synthesized Variables

1) Current: V3.0	2) Current Add MBA initial	3) Current Add TOB initial	4) Drop 6 Current Add MBA,TOB initial	5) Drop 4 Current Add TOB initial	6) Drop 6 Current Add TOB initial	7) Drop 6 Current Add TOB initial and 2000
black	black	black	-	-	-	-
male	male	male	male	male	male	male
educ_3cat	educ_3cat	educ_3cat	-	educ_3cat	-	-
maritalstat	maritalstat	maritalstat	maritalstat	maritalstat	maritalstat	maritalstat
age_cat	age_cat	age_cat	-	-	-	-
black_spouse	black_spouse	black_spouse	-	-	-	-
male_spouse	male_spouse	male_spouse	male_spouse	male_spouse	male_spouse	male_spouse
educ_3cat_spouse	educ_3cat_spouse	educ_3cat_spouse	-	educ_3cat_spouse	-	-
age_cat_spouse	age_cat_spouse	age_cat_spouse	-	-	-	-
-	MBA initial	-	MBA initial	-	-	-
-	MBA initial spouse	-	MBA initial spouse	-	-	-
-	-	tob_initial	tob_initial	tob_initial	tob_initial	tob_initial
-	-	tob_initial_spouse	tob_initial_spouse	tob_initial_spouse	tob_initial_spouse	tob_initial_spouse
-	-	-	-	-	-	tob_2000
-	-	-	-	-	-	tob_2000_spouse
SMALL CELL ANALYSIS AT PERSON-LEVEL						
Number of small person-level cells (<10 individuals):						
DBV: 159	DBV: 16048	DBV: 1952	DBV: 8092	DBV: 196	DBV: 26	DBV: 129
EBV: 159	EBV: 17240	EBV: 1967	EBV: 7952	EBV: 194	EBV: 26	EBV: 133
Number of individuals in small person-level cells:						
DBV: 472	DBV: 37611	DBV: 5448	DBV: 20251	DBV: 672	DBV: 76	DBV: 373
EBV: 472	EBV: 38191	EBV: 5387	EBV: 19920	EBV: 648	EBV: 72	EBV: 373
SMALL CELL ANALYSIS AT HOUSEHOLD-LEVEL						
Number of small household-level cells (<10 households):						
DBV: 93	DBV: 8724	DBV: 1074	DBV: 4238	DBV: 113	DBV: 19	DBV: 85
EBV: 93	EBV: 9333	EBV: 1081	EBV: 4171	EBV: 112	EBV: 19	EBV: 86
Number of households in small household-level cells:						
DBV: 280	DBV: 21143	DBV: 3048	DBV: 10807	DBV: 379	DBV: 54	DBV: 246
EBV: 280	EBV: 21437	EBV: 3020	EBV: 10650	EBV: 367	EBV: 52	EBV: 245
MBA initial and MBA initial spouse are rounded to nearest \$50						
tob_initial, tob_initial spouse, tob_2000, and tob_2000 spouse are put in categories of 1,2,3,5, and other						

Table 2: Analytic Validity of SIPP-PUF Weights -- Weighted Counts of Benefit Recipients

	SSA Reports	Average across completed data using completed weights	Average across synthetic data using synthetic weights	Percentage difference between columns C and D
Number of retired workers receiving benefits in Dec. 2000 (in millions)	28.50	27.10	26.40	-2.58
Average monthly benefit for retired workers	845.00	820.00	824.00	0.49
Number of widows and widowers receiving benefits in Dec. 2000 (in millions)	4.70	4.50	4.30	-4.44
Average monthly benefit for widows and widowers	810.00	752.00	753.00	0.13
Number of disabled receiving benefits in Dec. 2000 (in millions)	5.00	5.90	5.90	0.00
Average monthly benefit for disabled	786.00	736.00	738.00	0.27
Number of permanently insured individuals in Dec. 2000 (in millions)	140.70	131.40	133.70	1.75
DER average earnings for 2000	31,213.00	33,331.00	34,751.00	4.26
Number of wage and salary workers w/taxable earnings for 2000 (in millions)	145.00	128.00	129.00	0.78
SER average earnings for 2000	26,081.00	27,360.00	28,196.00	3.06

Table 3: SER work indicator for year 1965

Demographic Group	Type of Benefit	Mean		Confidence Interval		Confidence Interval		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	0.567	0.563	0.562	0.572	0.557	0.569	0	0.00001	0.00001
	disability	0.374	0.374	0.361	0.386	0.360	0.388	0	0.00005	0.00007
	aged spouse	0.133	0.130	0.122	0.144	0.114	0.147	0	0.00004	0.00008
	aged widow	0.225	0.210	0.216	0.233	0.200	0.219	0	0.00002	0.00003
	other	0.057	0.052	0.053	0.062	0.048	0.056	0	0.00001	0.00001
black females	own retirement	0.658	0.693	0.642	0.674	0.668	0.717	0	0.00009	0.00019
	disability	0.347	0.317	0.316	0.378	0.291	0.344	0	0.00030	0.00024
	aged spouse	0.255	0.225	0.209	0.301	0.185	0.264	0	0.00066	0.00057
	aged widow	0.300	0.359	0.262	0.339	0.313	0.405	0	0.00047	0.00069
	other	0.044	0.046	0.034	0.054	0.036	0.056	0	0.00003	0.00003
white males	own retirement	0.895	0.895	0.889	0.900	0.891	0.898	0	0.00001	0.00000
	disability	0.563	0.556	0.548	0.579	0.542	0.570	0	0.00007	0.00007
	aged spouse	0.161	0.166	0.070	0.253	0.093	0.240	0	0.00260	0.00191
	aged widow	0.414	0.456	0.292	0.535	0.334	0.579	0	0.00510	0.00550
	other	0.009	0.007	0.007	0.011	0.005	0.009	0	0.00000	0.00000
black males	own retirement	0.858	0.865	0.847	0.868	0.852	0.879	0	0.00004	0.00007
	disability	0.465	0.448	0.435	0.496	0.417	0.479	0	0.00028	0.00032
	aged spouse	0.202	0.079	0.068	0.335	-0.027	0.184	0	0.00291	0.00408
	aged widow	0.296	0.253	0.110	0.483	0.051	0.456	0	0.01189	0.01462
	other	0.006	0.003	0.002	0.009	0.000	0.005	0	0.00000	0.00000

Table 4: SER earnings in year 1965

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	1,461.89	1,523.24	1,456.39	1,467.40	1,496.23	1,550.25	1	10.49	252.06
	disability	701.12	731.24	675.21	727.04	693.82	768.65	0	242.88	484.23
	aged spouse	148.21	164.41	123.88	172.55	139.32	189.51	0	175.13	204.31
	aged widow	350.18	323.48	339.50	360.85	303.76	343.20	1	39.45	143.18
	other	70.24	71.31	62.63	77.85	62.39	80.24	0	20.30	28.00
black females	own retirement	1,370.90	1,522.66	1,234.12	1,507.69	1,442.74	1,602.57	0	3,937.78	2,091.20
	disability	486.02	483.09	410.19	561.85	435.22	530.96	0	1,287.51	837.98
	aged spouse	279.02	296.24	166.64	391.41	228.57	363.91	0	3,385.90	1,691.88
	aged widow	373.80	451.48	278.31	469.28	348.49	554.48	0	1,953.18	3,319.55
	other	40.04	37.21	27.44	52.65	25.58	48.84	0	53.31	47.73
white males	own retirement	3,678.21	3,840.81	3,655.68	3,700.75	3,821.64	3,859.97	0	122.39	135.54
	disability	1,815.13	1,843.00	1,764.99	1,865.28	1,801.70	1,884.29	0	796.70	627.67
	aged spouse	364.09	462.62	128.70	599.49	214.68	710.56	0	17,818.49	21,517.48
	aged widow	1,075.53	745.42	694.18	1,456.89	194.92	1,295.91	0	52,089.95	93,876.78
	other	8.78	8.77	5.09	12.46	5.57	11.98	0	4.63	3.74
black males	own retirement	3,019.57	3,067.00	2,912.03	3,127.10	2,982.02	3,151.98	1	4,001.08	2,491.36
	disability	1,254.48	1,205.29	1,113.37	1,395.59	1,066.37	1,344.21	0	3,927.03	5,661.00
	aged spouse	341.05	115.63	26.88	655.23	-38.88	270.15	0	23,592.61	8,822.80
	aged widow	698.37	422.35	-361.61	1,758.35	15.54	829.16	0	332,083.19	59,957.88
	other	3.45	1.25	-0.02	6.92	-1.44	3.93	0	4.41	2.64

Table 5: SER work indicator for year 1975

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	0.675	0.690	0.666	0.684	0.682	0.698	0	0.00002	0.00002
	disability	0.554	0.548	0.532	0.576	0.530	0.566	0	0.00012	0.00010
	aged spouse	0.158	0.140	0.148	0.167	0.130	0.151	0	0.00003	0.00004
	aged widow	0.250	0.242	0.241	0.259	0.225	0.260	0	0.00003	0.00008
	other	0.217	0.211	0.208	0.225	0.204	0.219	0	0.00003	0.00002
black females	own retirement	0.722	0.742	0.704	0.739	0.722	0.762	0	0.00008	0.00014
	disability	0.573	0.599	0.554	0.591	0.577	0.621	0	0.00012	0.00018
	aged spouse	0.268	0.297	0.231	0.306	0.248	0.346	0	0.00047	0.00086
	aged widow	0.303	0.303	0.273	0.333	0.255	0.352	0	0.00032	0.00074
	other	0.179	0.189	0.166	0.191	0.172	0.205	0	0.00005	0.00010
white males	own retirement	0.865	0.879	0.860	0.870	0.875	0.883	0	0.00001	0.00001
	disability	0.705	0.706	0.696	0.713	0.697	0.716	0	0.00003	0.00003
	aged spouse	0.135	0.106	0.076	0.193	0.051	0.160	0	0.00124	0.00109
	aged widow	0.366	0.392	0.238	0.494	0.244	0.539	0	0.00548	0.00733
	other	0.194	0.191	0.183	0.205	0.182	0.201	0	0.00004	0.00003
black males	own retirement	0.790	0.814	0.764	0.815	0.794	0.833	0	0.00015	0.00013
	disability	0.648	0.661	0.626	0.669	0.636	0.686	0	0.00015	0.00022
	aged spouse	0.136	0.309	-0.042	0.314	-0.139	0.757	0	0.00841	0.04861
	aged widow	0.252	0.000	0.035	0.469			0	0.01382	0.00000
	other	0.159	0.147	0.130	0.187	0.128	0.165	0	0.00020	0.00012

Table 6: SER Earnings for year 1975

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	3,948	4,144	3,655	4,242	4,078	4,211	0	10,404	1,493
	disability	2,274	2,279	2,137	2,411	2,168	2,391	0	4,464	3,918
	aged spouse	359	315	300	418	280	350	0	1,016	451
	aged widow	854	830	798	911	760	900	0	1,022	1,494
	other	543	541	462	625	511	571	0	1,245	327
black females	own retirement	3,878	4,308	3,635	4,121	4,158	4,458	0	13,942	8,311
	disability	2,211	2,435	1,970	2,452	2,287	2,582	0	14,500	7,967
	aged spouse	581	628	404	759	462	794	0	10,334	9,718
	aged widow	1,033	1,040	889	1,177	814	1,265	0	7,260	16,326
	other	419	459	317	520	392	526	0	2,787	1,578
white males	own retirement	9,241	9,836	9,197	9,285	9,770	9,902	1	663	1,564
	disability	5,446	5,574	5,329	5,563	5,465	5,683	1	4,730	4,371
	aged spouse	340	180	131	548	72	288	0	15,455	4,330
	aged widow	3,032	2,644	1,640	4,425	1,161	4,127	0	643,749	747,378
	other	630	589	570	690	546	631	0	586	661
black males	own retirement	6,990	7,500	6,782	7,199	7,296	7,704	1	15,086	15,303
	disability	3,934	4,037	3,691	4,176	3,768	4,306	0	18,273	24,563
	aged spouse	401	164	-263	1,064	-92	420	0	117,515	18,317
	aged widow	1,602	0	-970	4,174			0	2,061,954	0
	other	367	339	262	471	271	407	0	3,056	1,648

Table 7: SER work indicator for year 1985

Demographic Group	Type of Benefit	Mean		Confidence Interval		Confidence Interval		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	0.547	0.557	0.531	0.563	0.539	0.574	0	0.00006	0.00007
	disability	0.594	0.597	0.574	0.614	0.579	0.615	0	0.00011	0.00010
	aged spouse	0.150	0.159	0.116	0.183	0.129	0.189	0	0.00025	0.00021
	aged widow	0.210	0.209	0.194	0.226	0.192	0.226	0	0.00007	0.00008
	other	0.419	0.429	0.407	0.432	0.418	0.440	0	0.00005	0.00004
black females	own retirement	0.572	0.558	0.524	0.619	0.533	0.583	0	0.00051	0.00020
	disability	0.605	0.593	0.568	0.641	0.544	0.643	0	0.00037	0.00064
	aged spouse	0.165	0.133	0.118	0.213	0.082	0.183	0	0.00068	0.00079
	aged widow	0.224	0.222	0.189	0.259	0.190	0.254	0	0.00040	0.00037
	other	0.366	0.353	0.344	0.387	0.334	0.372	0	0.00016	0.00014
white males	own retirement	0.673	0.679	0.660	0.686	0.668	0.690	0	0.00004	0.00003
	disability	0.634	0.640	0.621	0.646	0.625	0.656	0	0.00005	0.00007
	aged spouse	0.189	0.206	0.104	0.274	0.137	0.275	0	0.00245	0.00177
	aged widow	0.301	0.293	0.151	0.451	0.150	0.437	0	0.00704	0.00680
	other	0.459	0.466	0.441	0.478	0.452	0.480	0	0.00009	0.00007
black males	own retirement	0.630	0.607	0.597	0.663	0.556	0.658	0	0.00023	0.00062
	disability	0.593	0.572	0.566	0.621	0.548	0.595	0	0.00024	0.00020
	aged spouse	0.075	0.071	-0.040	0.189	-0.081	0.222	0	0.00438	0.00713
	aged widow	0.250	0.101	0.051	0.450	-0.023	0.226	0	0.01399	0.00570
	other	0.380	0.387	0.346	0.413	0.358	0.417	0	0.00036	0.00030

Table 8: SER Earnings for year 1985

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	7,101	7,200	6,945	7,258	6,927	7,472	1	8,438	18,878
	disability	6,069	5,859	5,636	6,502	5,673	6,045	0	30,118	12,792
	aged spouse	960	892	565	1,355	501	1,282	0	29,660	31,332
	aged widow	1,842	1,724	1,661	2,023	1,510	1,938	0	8,060	12,058
	other	3,656	3,582	3,527	3,784	3,375	3,790	1	5,725	12,547
black females	own retirement	6,915	7,057	6,116	7,714	6,657	7,457	0	131,457	55,966
	disability	5,655	5,796	5,311	6,000	5,277	6,314	0	31,031	84,628
	aged spouse	985	903	638	1,333	457	1,348	0	39,258	60,710
	aged widow	1,694	1,692	1,391	1,996	1,349	2,035	0	32,491	43,063
	other	2,665	2,576	2,434	2,897	2,352	2,800	0	17,039	18,195
white males	own retirement	15,254	16,091	15,007	15,500	15,705	16,477	1	20,992	39,152
	disability	10,186	10,573	9,655	10,718	10,140	11,005	0	67,302	54,135
	aged spouse	1,385	1,420	501	2,269	506	2,334	0	266,884	275,395
	aged widow	4,318	4,981	1,339	7,296	1,119	8,843	0	2,880,355	4,575,582
	other	5,966	6,003	5,724	6,208	5,742	6,264	0	16,686	23,163
black males	own retirement	11,381	11,409	10,926	11,836	10,155	12,664	1	71,721	386,020
	disability	7,870	7,563	7,127	8,613	7,093	8,033	0	140,174	81,435
	aged spouse	368	907	-387	1,123	-656	2,470	0	206,113	871,879
	aged widow	1,647	524	-755	4,049	-176	1,223	0	1,728,091	178,634
	other	3,821	3,582	3,360	4,283	3,194	3,970	0	69,510	53,807

Table 9: SER work indicator for year 1995

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	0.338	0.343	0.329	0.347	0.336	0.349	0	0.00002	0.00002
	disability	0.452	0.452	0.441	0.463	0.440	0.463	0	0.00004	0.00005
	aged spouse	0.109	0.112	0.079	0.140	0.080	0.144	0	0.00020	0.00022
	aged widow	0.157	0.154	0.149	0.165	0.145	0.162	0	0.00002	0.00003
	other	0.606	0.613	0.593	0.619	0.598	0.627	0	0.00005	0.00006
black females	own retirement	0.373	0.362	0.356	0.391	0.343	0.382	0	0.00010	0.00014
	disability	0.460	0.456	0.417	0.503	0.429	0.482	0	0.00046	0.00025
	aged spouse	0.122	0.107	0.060	0.183	0.066	0.147	0	0.00097	0.00054
	aged widow	0.173	0.186	0.138	0.207	0.148	0.223	0	0.00036	0.00045
	other	0.597	0.576	0.564	0.631	0.552	0.600	0	0.00029	0.00020
white males	own retirement	0.422	0.444	0.413	0.431	0.436	0.451	0	0.00002	0.00002
	disability	0.439	0.457	0.409	0.469	0.434	0.481	0	0.00020	0.00014
	aged spouse	0.095	0.107	-0.006	0.196	0.023	0.190	0	0.00267	0.00214
	aged widow	0.273	0.284	0.179	0.366	0.164	0.403	0	0.00320	0.00507
	other	0.701	0.691	0.678	0.724	0.670	0.712	0	0.00013	0.00012
black males	own retirement	0.435	0.413	0.388	0.481	0.393	0.433	0	0.00032	0.00015
	disability	0.443	0.450	0.422	0.463	0.425	0.475	0	0.00015	0.00023
	aged spouse	0.159	0.233	-0.094	0.412	-0.181	0.646	0	0.01659	0.04149
	aged widow	0.147	0.235	-0.016	0.310	0.052	0.418	1	0.00919	0.01221
	other	0.628	0.618	0.597	0.659	0.584	0.652	0	0.00013	0.00037

Table 10: SER Earnings for year 1995

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	5,495	5,566	5,423	5,566	5,421	5,710	1	1,764	7,686
	disability	6,135	6,084	5,911	6,359	5,757	6,411	0	12,002	35,478
	aged spouse	1,106	1,095	678	1,533	466	1,724	0	40,062	80,651
	aged widow	1,849	1,760	1,732	1,966	1,566	1,954	0	4,632	11,808
	other	9,044	8,969	8,808	9,281	8,597	9,342	1	19,351	42,286
black females	own retirement	5,630	5,501	5,183	6,078	5,028	5,973	0	55,905	79,479
	disability	5,590	6,032	4,840	6,341	5,383	6,681	0	150,415	137,388
	aged spouse	1,258	1,108	428	2,088	605	1,611	0	186,677	87,164
	aged widow	2,003	1,975	1,526	2,480	1,488	2,463	0	78,477	87,268
	other	7,341	6,706	6,903	7,779	6,275	7,138	0	66,509	67,843
white males	own retirement	10,347	11,012	10,003	10,691	10,751	11,274	0	21,142	23,637
	disability	8,756	9,173	7,735	9,776	8,271	10,075	0	213,915	187,909
	aged spouse	811	951	-520	2,142	-427	2,328	0	480,470	520,068
	aged widow	3,971	3,855	1,017	6,924	1,170	6,540	0	3,007,166	2,639,770
	other	15,624	15,661	14,583	16,664	15,105	16,216	0	224,431	98,502
black males	own retirement	8,856	8,564	7,853	9,858	7,883	9,245	1	347,683	163,956
	disability	6,569	5,915	5,751	7,387	5,388	6,441	0	175,129	101,723
	aged spouse	1,691	3,838	-2,110	5,492	-5,596	13,272	0	3,074,159	20,071,412
	aged widow	1,012	730	-1,743	3,767	-2,054	3,513	0	2,685,640	1,947,740
	other	9,757	9,309	9,093	10,421	8,614	10,003	0	153,850	176,484

Table 11: Total SER earnings 1951-2003

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	192,468	198,303	189,034	195,902	195,018	201,589	0	3,635,902	3,582,412
	disability	159,975	160,721	155,261	164,689	153,560	167,882	0	7,290,556	14,186,930
	aged spouse	31,945	32,601	20,290	43,600	18,207	46,996	0	27,166,130	40,572,680
	aged widow	52,794	51,821	49,392	56,195	47,184	56,459	0	3,602,536	5,853,323
	other	187,463	187,956	182,067	192,860	180,912	194,999	0	9,256,167	14,299,967
black females	own retirement	191,274	195,617	180,979	201,570	187,629	203,605	0	27,671,924	23,120,206
	disability	145,353	151,265	137,949	152,757	140,394	162,136	0	18,055,260	35,532,660
	aged spouse	36,723	36,296	25,665	47,782	25,237	47,356	0	36,155,536	39,974,037
	aged widow	56,721	57,379	48,964	64,478	48,166	66,592	0	22,018,177	31,222,865
	other	152,606	146,146	144,237	160,975	138,596	153,697	0	23,731,960	20,555,483
white males	own retirement	417,976	442,503	413,684	422,268	438,563	446,442	0	5,837,279	5,662,728
	disability	276,091	288,266	254,564	297,618	268,521	308,011	0	94,001,775	82,880,905
	aged spouse	33,447	32,596	9,986	56,908	12,442	52,749	0	143,099,719	111,939,873
	aged widow	126,429	134,014	67,633	185,225	71,111	196,917	0	1,227,132,456	1,441,472,756
	other	315,302	319,194	300,010	330,593	306,393	331,994	0	59,030,061	45,981,602
black males	own retirement	330,958	331,280	311,262	350,654	317,708	344,852	1	134,237,271	63,661,959
	disability	204,902	197,208	186,983	222,821	185,899	208,516	0	82,954,046	43,726,354
	aged spouse	48,022	66,377	-25,930	121,974	-46,152	178,906	0	1,289,053,431	2,882,382,428
	aged widow	56,265	29,515	5,535	106,994	-2,984	62,014	0	841,272,477	309,581,707
	other	200,003	194,732	186,450	213,556	182,929	206,534	0	63,536,541	51,252,661

Table 12: Total years worked in SER (i.e. positive FICA earnings)

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	26.174	26.693	25.881	26.466	26.448	26.939	0	0.02243	0.01731
	disability	21.679	22.076	21.387	21.972	21.776	22.376	0	0.02815	0.02983
	aged spouse	8.047	8.099	7.189	8.904	7.172	9.026	0	0.15682	0.18089
	aged widow	10.614	10.353	10.349	10.879	10.096	10.609	0	0.02540	0.02425
	other	15.050	15.459	14.771	15.328	15.208	15.710	0	0.02286	0.01998
black females	own retirement	27.847	28.428	26.900	28.794	27.931	28.925	0	0.21083	0.08429
	disability	20.915	21.431	20.094	21.735	20.653	22.208	0	0.17740	0.17340
	aged spouse	10.317	9.974	8.972	11.663	8.792	11.156	0	0.55231	0.47391
	aged widow	12.594	13.320	11.495	13.694	12.293	14.346	0	0.38192	0.36726
	other	13.800	13.947	13.466	14.134	13.532	14.362	0	0.04064	0.06085
white males	own retirement	35.779	36.477	35.638	35.920	36.346	36.609	0	0.00654	0.00632
	disability	26.184	26.610	25.517	26.851	26.094	27.127	0	0.10068	0.06774
	aged spouse	8.108	8.506	6.575	9.641	6.615	10.398	0	0.86016	1.26191
	aged widow	15.958	15.778	11.908	20.008	11.962	19.593	0	5.75307	5.37211
	other	15.907	16.243	15.403	16.410	15.844	16.642	0	0.06124	0.04372
black males	own retirement	33.902	33.791	33.230	34.574	33.284	34.298	1	0.15613	0.09151
	disability	23.579	23.571	23.040	24.119	22.771	24.371	0	0.10190	0.19847
	aged spouse	10.429	9.296	7.422	13.437	4.335	14.257	0	1.19750	5.85639
	aged widow	14.820	10.428	7.940	21.701	6.120	14.735	0	15.52250	6.22151
	other	13.783	13.979	13.202	14.365	13.505	14.452	0	0.11294	0.08209

Table 13: Personal Account: 2% of earnings compounded annually at 5% interest from 1951 until date of initial entitlement

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	7,177	7,532	6,975	7,379	7,348	7,715	0	9,859	9,142
	disability	4,976	5,140	4,774	5,179	4,993	5,287	0	11,836	7,725
	aged spouse	702	692	412	991	366	1,018	0	15,763	20,017
	aged widow	1,726	1,710	1,643	1,808	1,615	1,805	0	2,455	3,175
	other	1,187	1,242	1,113	1,261	1,150	1,334	0	1,995	2,966
black females	own retirement	7,247	7,656	6,871	7,623	7,356	7,956	0	33,289	32,941
	disability	4,465	4,849	4,261	4,670	4,531	5,167	0	14,657	33,688
	aged spouse	707	664	475	938	408	919	0	13,710	17,379
	aged widow	2,038	2,256	1,746	2,330	1,857	2,656	0	31,310	57,392
	other	1,139	1,282	944	1,334	1,087	1,477	0	13,535	14,062
white males	own retirement	16,789	17,985	16,505	17,074	17,743	18,227	0	17,721	17,344
	disability	9,321	9,945	9,023	9,618	9,724	10,167	0	25,702	17,573
	aged spouse	1,284	1,401	643	1,925	663	2,138	0	115,566	146,144
	aged widow	5,767	6,209	3,201	8,333	3,324	9,095	0	2,314,225	3,036,071
	other	1,529	1,802	1,160	1,898	1,382	2,222	0	49,990	65,210
black males	own retirement	13,730	13,975	12,802	14,658	13,305	14,646	1	298,247	139,161
	disability	6,835	6,803	5,811	7,859	6,426	7,180	0	233,733	51,875
	aged spouse	1,436	1,187	1,024	1,848	48	2,326	1	58,685	284,640
	aged widow	2,495	1,240	-106	5,096	485	1,996	0	2,180,277	204,410
	other	1,409	1,883	771	2,047	869	2,897	0	148,969	370,626

Table 14: Year of initial entitlement

Demographic Group	Type of Benefit	Mean		Confidence Interval		Confidence Interval		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	1990.495	1990.365	1990.366	1990.625	1990.233	1990.497	0	0.00521	0.00584
	disability	1991.684	1991.793	1991.491	1991.877	1991.537	1992.049	0	0.01357	0.02242
	aged spouse	1989.963	1989.739	1989.637	1990.290	1989.352	1990.126	0	0.03601	0.04880
	aged widow	1985.799	1985.683	1985.596	1986.002	1985.442	1985.924	0	0.01483	0.01989
	other	1980.468	1980.537	1980.201	1980.735	1980.334	1980.741	0	0.02160	0.01468
black females	own retirement	1990.550	1990.641	1990.197	1990.903	1990.141	1991.141	0	0.02547	0.07539
	disability	1992.005	1992.477	1991.659	1992.351	1991.923	1993.031	0	0.04377	0.09590
	aged spouse	1985.726	1984.274	1984.200	1987.252	1982.849	1985.698	0	0.75069	0.73582
	aged widow	1986.729	1987.628	1986.062	1987.397	1986.814	1988.442	0	0.15422	0.23381
	other	1980.272	1980.228	1979.910	1980.634	1979.721	1980.735	0	0.04777	0.08356
white males	own retirement	1991.261	1991.415	1991.114	1991.409	1991.317	1991.513	0	0.00318	0.00340
	disability	1990.464	1990.757	1990.189	1990.738	1990.514	1991.001	0	0.02214	0.01963
	aged spouse	1991.398	1991.777	1989.431	1993.365	1989.445	1994.108	0	1.17039	1.56484
	aged widow	1990.582	1991.390	1987.960	1993.204	1988.920	1993.861	0	2.23807	2.20803
	other	1978.964	1979.042	1978.706	1979.222	1978.812	1979.272	0	0.02174	0.01876
black males	own retirement	1992.160	1992.315	1991.870	1992.450	1992.003	1992.627	1	0.02911	0.03499
	disability	1990.428	1990.793	1989.913	1990.944	1990.335	1991.250	0	0.08794	0.07576
	aged spouse	1993.600	1994.522	1991.709	1995.492	1992.364	1996.680	0	1.01994	1.50382
	aged widow	1988.757	1988.541	1984.973	1992.541	1985.856	1991.227	0	4.72418	2.58058
	other	1979.383	1979.345	1978.910	1979.856	1978.894	1979.796	0	0.07976	0.07508

Table 15: Initial Monthly Benefit Amount

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	441.09	444.99	437.29	444.89	441.19	448.79	0	4.88	5.05
	disability	498.25	495.18	490.27	506.24	485.83	504.53	0	21.96	28.29
	aged spouse	330.97	337.90	325.31	336.62	332.40	343.40	0	11.04	10.63
	aged widow	384.91	388.70	374.64	395.17	377.27	400.13	0	31.14	37.17
	other	169.50	171.87	162.97	176.02	165.89	177.85	0	11.27	9.97
black females	own retirement	430.14	439.66	416.91	443.36	427.77	451.54	0	51.22	49.65
	disability	482.48	495.58	467.29	497.68	474.45	516.71	0	75.18	130.05
	aged spouse	255.94	241.41	230.75	281.14	223.13	259.70	0	190.50	122.32
	aged widow	381.71	389.41	358.37	405.05	357.10	421.72	0	169.10	324.52
	other	136.92	130.13	127.00	146.84	120.37	139.90	0	28.09	27.63
white males	own retirement	716.69	730.33	710.73	722.66	725.39	735.28	0	9.39	8.31
	disability	680.11	685.26	672.84	687.39	676.86	693.65	0	19.34	25.20
	aged spouse	236.14	238.09	210.16	262.13	212.11	264.06	0	246.03	246.42
	aged widow	449.02	457.16	311.18	586.86	326.82	587.49	0	5,745.09	5,608.83
	other	156.91	158.97	153.71	160.12	155.47	162.46	0	3.75	4.44
black males	own retirement	630.27	629.62	606.77	653.78	615.87	643.37	1	191.17	68.41
	disability	598.74	589.93	572.92	624.57	575.24	604.61	0	182.97	79.67
	aged spouse	278.44	319.22	208.73	348.15	206.15	432.29	0	1,646.79	3,656.73
	aged widow	334.20	331.05	203.67	464.74	191.72	470.38	0	5,382.22	5,362.25
	other	129.48	123.13	122.55	136.41	116.54	129.72	0	16.31	16.01

Table 16: Monthly Benefit Amount April 2000

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	583	563	573	594	558	569	0	27	12
	disability	581	598	564	598	591	605	0	59	16
	aged spouse	560	598	551	569	590	606	0	24	22
	aged widow	542	584	530	554	567	602	0	48	98
	other	594	643	581	607	626	660	0	61	109
black females	own retirement	485	469	472	498	457	481	0	54	53
	disability	442	445	429	456	431	459	0	55	69
	aged spouse	430	448	400	460	418	477	0	229	254
	aged widow	444	450	414	474	408	492	0	305	634
	other	507	620	415	598	558	682	0	1,891	1,313
white males	own retirement	715	709	682	749	700	717	0	214	25
	disability	719	739	685	753	731	747	0	205	23
	aged spouse	708	745	682	735	734	755	0	138	40
	aged widow	786	812	758	814	796	828	0	200	91
	other	844	886	796	893	869	904	0	421	106
black males	own retirement	598	581	563	633	562	599	0	229	118
	disability	538	522	489	587	501	544	0	472	167
	aged spouse	514	490	487	540	462	518	0	228	276
	aged widow	567	584	519	616	490	678	0	837	2,682
	other	701	650	574	828	583	717	0	3,801	1,652

Table 17: Average Indexed Monthly Earnings or Average Monthly Wage

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	own retirement	725	760	707	743	744	776	0	80	71
	disability	896	907	877	916	883	932	0	138	204
	aged spouse	112	116	90	135	86	145	0	114	179
	aged widow	195	193	187	204	183	203	0	25	35
	other	1,184	1,180	1,157	1,210	1,149	1,210	0	249	313
black females	own retirement	724	773	693	756	746	801	0	248	276
	disability	833	870	784	881	812	927	0	670	986
	aged spouse	151	159	123	179	116	201	0	263	582
	aged widow	229	252	199	260	211	294	0	340	614
	other	978	949	914	1,041	903	996	0	1,269	800
white males	own retirement	1,693	1,789	1,666	1,720	1,771	1,807	0	161	103
	disability	1,621	1,691	1,546	1,695	1,619	1,762	0	1,237	1,196
	aged spouse	152	157	101	204	92	221	0	904	1,220
	aged widow	651	646	368	934	350	942	0	28,080	31,743
	other	2,016	2,031	1,954	2,078	1,969	2,093	0	1,156	1,198
black males	own retirement	1,373	1,399	1,293	1,453	1,348	1,450	1	2,227	866
	disability	1,244	1,213	1,153	1,336	1,159	1,267	0	2,184	1,026
	aged spouse	146	99	106	186	30	167	1	547	1,289
	aged widow	292	143	-29	613	55	231	0	32,478	2,800
	other	1,336	1,297	1,242	1,431	1,225	1,369	0	2,917	1,906

*AIME for individuals who reach age 62 after 1979, otherwise AMW

Table 18: Primary Insurance Amount

Demographic Group	Type of Benefit	Mean		Confidence Interval		Confidence Interval		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Completed	Completed		Synthetic	Completed
white females	own retirement	422	434	415	430	428	440	0	14	11
	disability	498	502	489	508	495	510	0	27	22
	aged spouse	100	103	81	118	80	126	0	70	104
	aged widow	178	181	170	185	173	188	0	17	19
	other	409	406	404	414	400	412	0	10	13
black females	own retirement	424	441	412	436	430	452	0	31	46
	disability	478	493	459	498	471	515	0	107	147
	aged spouse	118	113	96	139	88	137	0	134	169
	aged widow	184	197	166	203	169	224	0	124	250
	other	381	371	368	393	360	382	0	53	47
white males	own retirement	749	776	740	758	770	782	0	17	11
	disability	710	729	684	735	707	752	0	139	118
	aged spouse	149	162	108	190	113	211	0	544	742
	aged widow	395	366	261	530	209	523	0	6,414	8,824
	other	490	490	480	500	481	500	0	31	30
black males	own retirement	656	661	625	686	646	676	1	322	81
	disability	603	593	573	633	574	611	0	241	119
	aged spouse	138	110	109	168	57	162	1	303	820
	aged widow	199	123	33	365	45	201	0	9,143	2,171
	other	434	424	410	458	402	445	0	163	147

Table 19: SER work indicator for year 1965

Demographic Group	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white female	0.178	0.176	0.176	0.180	0.174	0.178	0	0.000001	0.000001
black female	0.143	0.148	0.139	0.148	0.143	0.154	0	0.000007	0.000010
white male	0.280	0.276	0.277	0.283	0.273	0.278	0	0.000003	0.000002
black male	0.212	0.199	0.203	0.221	0.190	0.207	0	0.000018	0.000023

Table 20: SER Earnings for year 1965

Demographic Group	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed						Synthetic	Completed
white female	354	375	347	362	369	381	1	18	12
black female	218	243	191	244	231	256	0	120	55
white male	916	940	905	927	931	950	0	35	30
black male	589	548	575	603	520	577	1	67	256

Table 21: SER work indicator for year 1975

Demographic Group	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white female	0.326	0.326	0.324	0.328	0.324	0.329	0	0.000001	0.000003
black female	0.285	0.301	0.278	0.292	0.294	0.307	0	0.000014	0.000015
white male	0.441	0.442	0.438	0.443	0.440	0.445	0	0.000003	0.000003
black male	0.356	0.349	0.343	0.369	0.340	0.357	0	0.000036	0.000026

Table 22: SER Earnings for year 1975

Demographic Group	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed						Synthetic	Completed
white female	1,382	1,457	1,255	1,509	1,441	1,472	0	2,410	89
black female	1,085	1,257	1,000	1,170	1,222	1,291	0	1,487	441
white male	3,533	3,668	3,486	3,581	3,631	3,705	1	771	446
black male	2,170	2,171	2,021	2,319	2,089	2,254	0	3,172	2,215

Table 23: SER work indicator for year 1985

Demographic Group	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white female	0.448	0.449	0.445	0.451	0.447	0.452	0	0.000003	0.000002
black female	0.400	0.397	0.391	0.409	0.391	0.404	0	0.000025	0.000015
white male	0.549	0.550	0.546	0.552	0.548	0.553	0	0.000003	0.000003
black male	0.453	0.434	0.445	0.460	0.426	0.441	0	0.000020	0.000021

Table 24: SER Earnings for year 1985

Demographic Group	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed						Synthetic	Completed
white female	4,804	4,723	4,680	4,929	4,684	4,761	1	5,388	552
black female	3,733	3,764	3,657	3,810	3,674	3,853	1	2,044	2,968
white male	9,892	10,091	9,763	10,021	10,023	10,160	1	5,746	1,728
black male	6,139	5,667	5,849	6,428	5,518	5,816	1	29,044	8,131

Table 25: SER work indicator for year 1995

Demographic Group	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white female	0.567	0.564	0.564	0.569	0.560	0.569	0	0.000002	0.000006
black female	0.570	0.545	0.555	0.586	0.536	0.554	0	0.000059	0.000025
white male	0.671	0.675	0.667	0.675	0.671	0.679	0	0.000005	0.000005
black male	0.606	0.581	0.598	0.615	0.573	0.590	0	0.000023	0.000024

Table 26: SER Earnings for year 1995

Demographic Group	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed						Synthetic	Completed
white female	9,561	9,443	9,441	9,680	9,365	9,522	1	4,964	2,193
black female	8,153	7,589	7,991	8,315	7,412	7,766	1	9,040	11,134
white male	17,047	17,241	16,837	17,257	17,114	17,367	1	15,300	5,512
black male	11,023	10,087	10,658	11,388	9,870	10,303	1	45,990	17,301

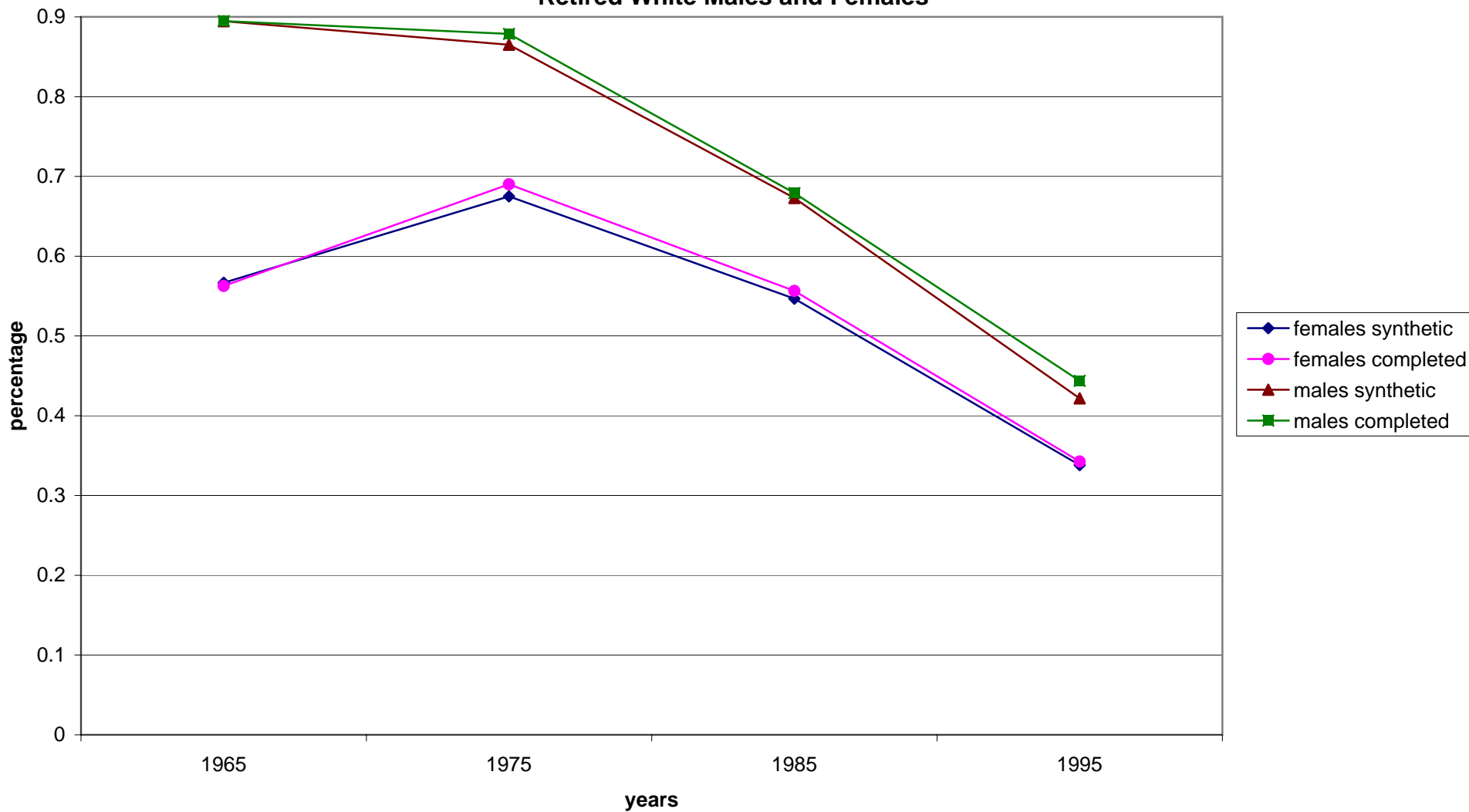
Table 27: Total SER Earnings 1951-2003

Demographic Group	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white female	211,817	212,751	210,419	213,214	211,508	213,995	0	489,904	564,662
black female	178,709	174,331	174,332	183,085	171,509	177,153	0	5,112,046	2,929,414
white male	400,702	409,869	398,904	402,500	407,744	411,994	0	795,526	1,607,808
black male	257,525	240,933	246,154	268,896	236,879	244,987	0	24,866,833	6,073,858

Table 28: Total Years worked in SER (i.e. positive FICA earnings)

Demographic Group	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed						Synthetic	Completed
white female	17.759	17.992	17.657	17.862	17.937	18.048	0	0.002876	0.001130
black female	16.310	16.569	16.128	16.491	16.421	16.717	0	0.010407	0.008026
white male	22.698	22.957	22.638	22.758	22.889	23.026	0	0.001320	0.001710
black male	18.992	18.412	18.607	19.376	18.184	18.640	0	0.023317	0.017944

Chart1:
Comparison of Synthetic and Completed Annual Work Indicators
Retired White Males and Females



**Chart 2:
Comparison of Synthetic and Completed Annual Work Indicators
Retired Black Males and Females**

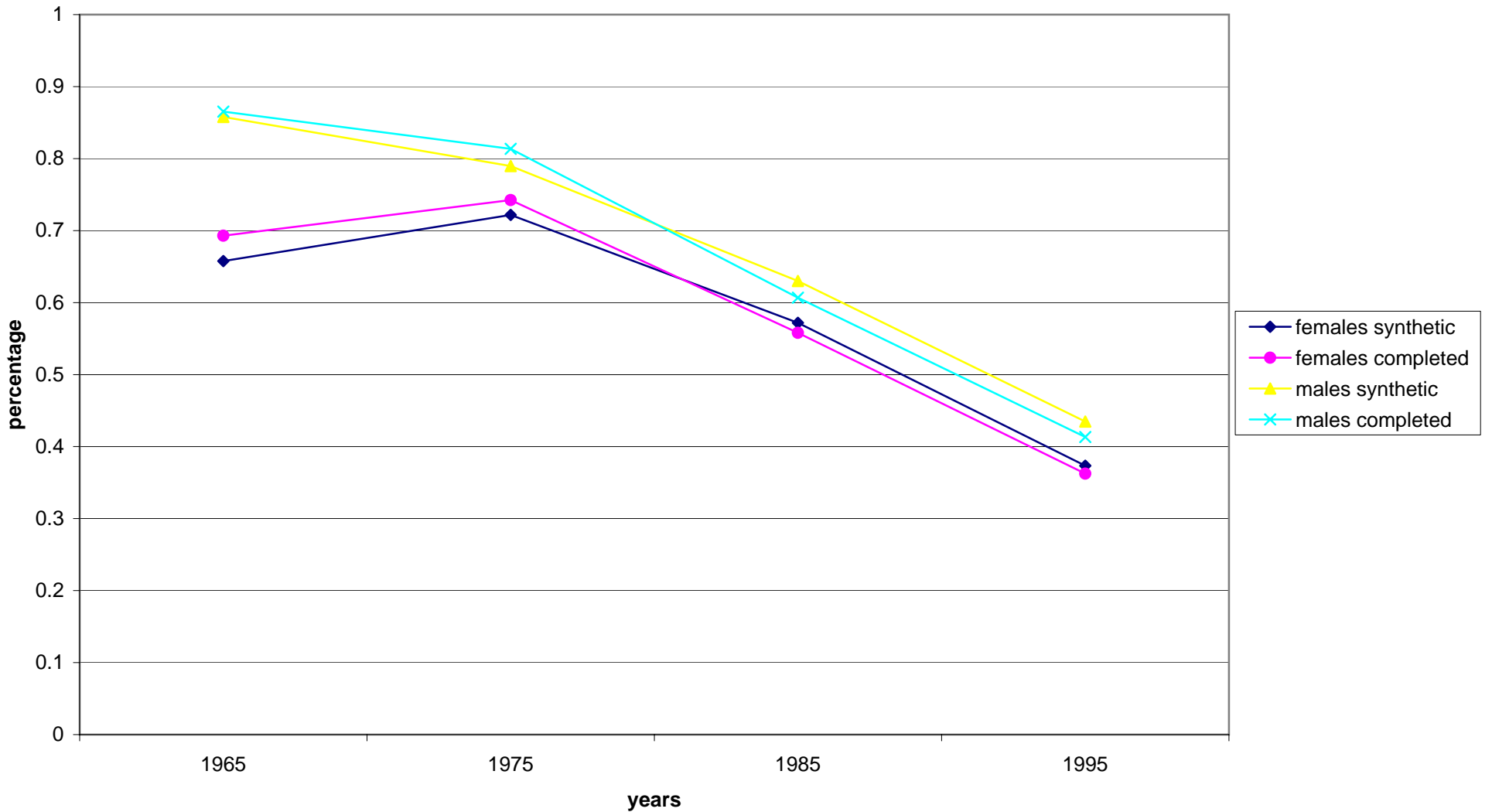


Chart 3:
Comparison of Synthetic and Completed Earnings
Retired White Males and Females

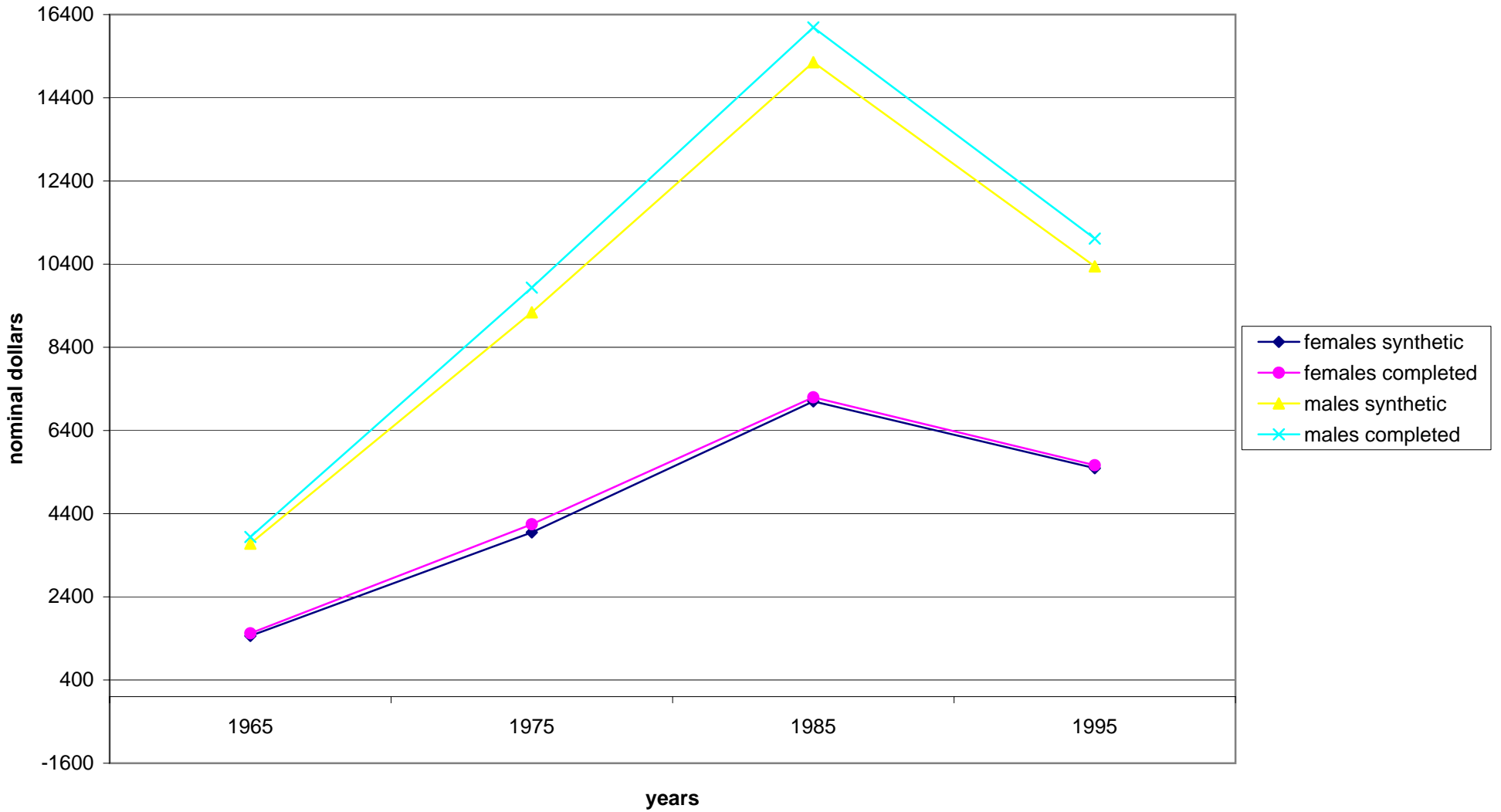


Chart 4:
Comparison of Synthetic and Completed Earnings
Retired Black Males and Females

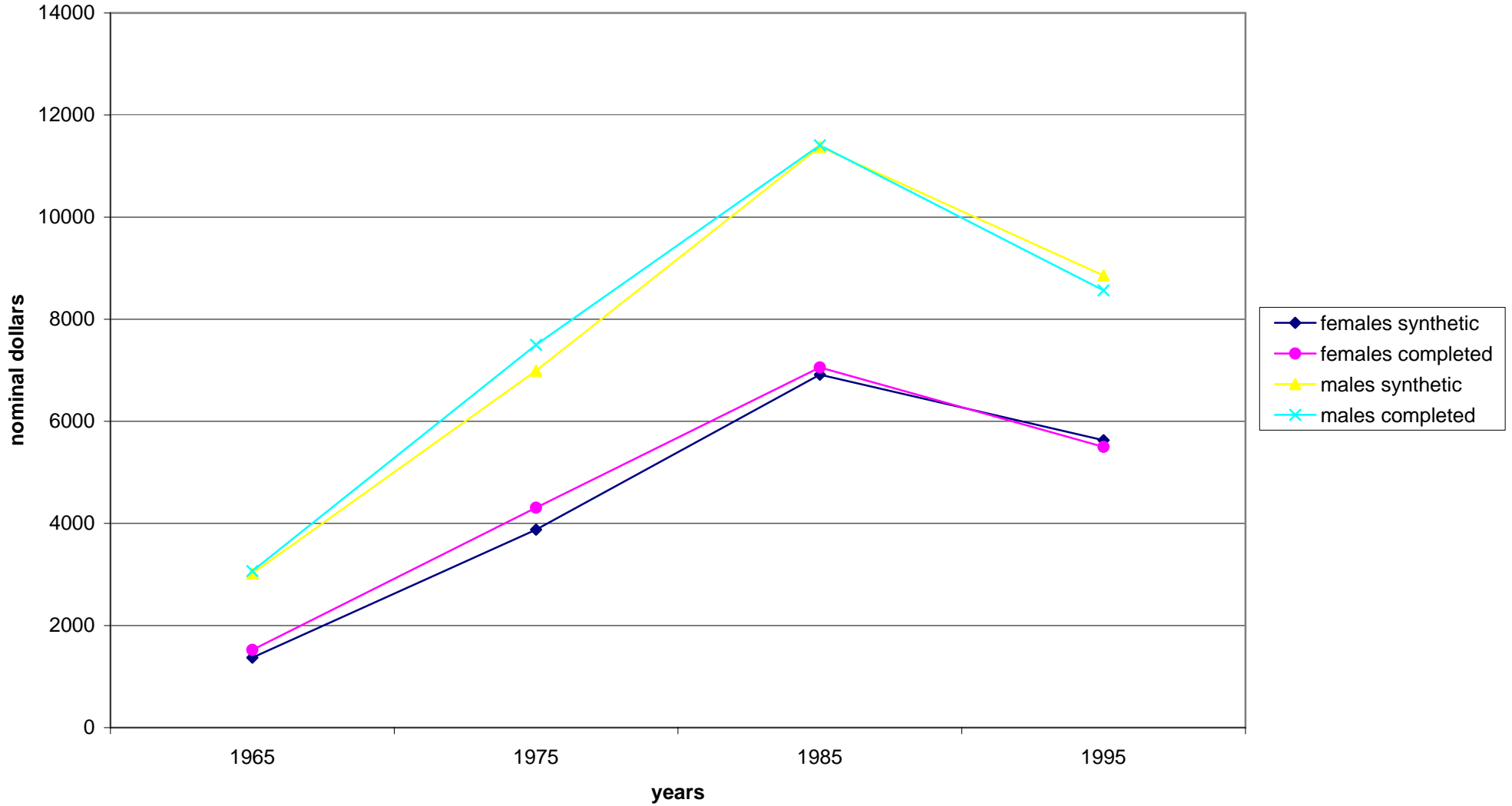


Chart 5:
Comparison of Synthetic and Completed Annual Work Indicators
White Males and Females

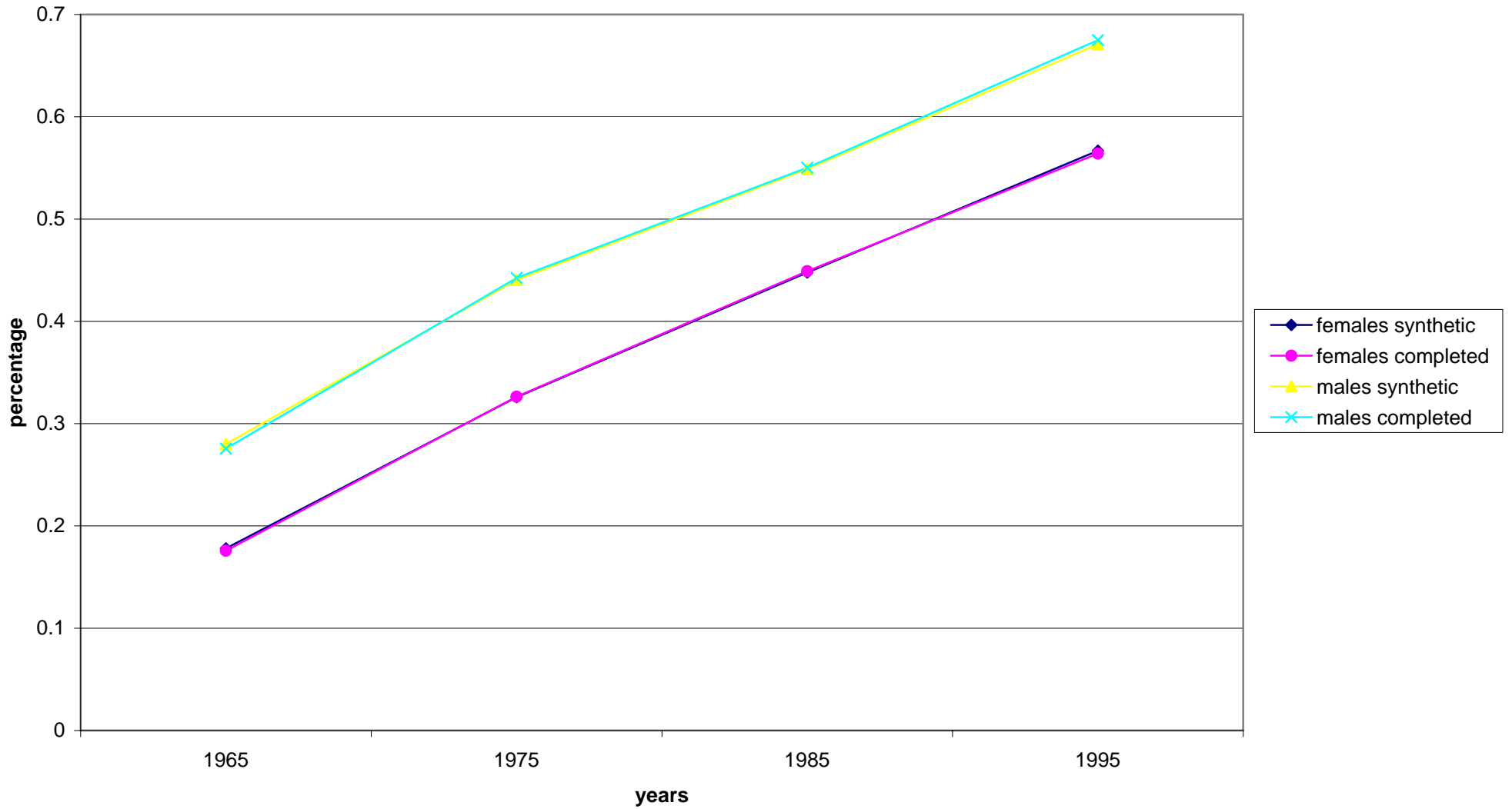
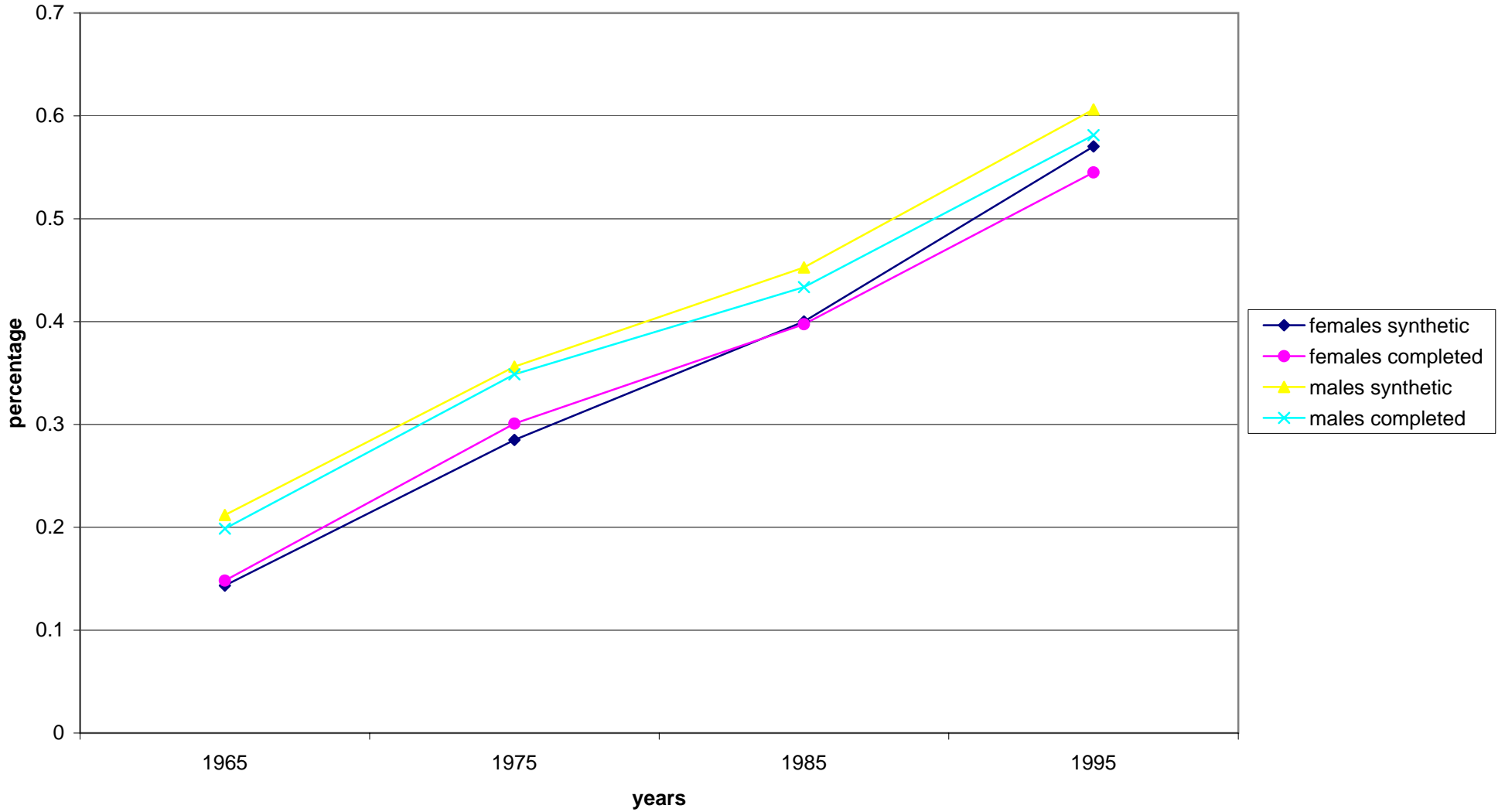
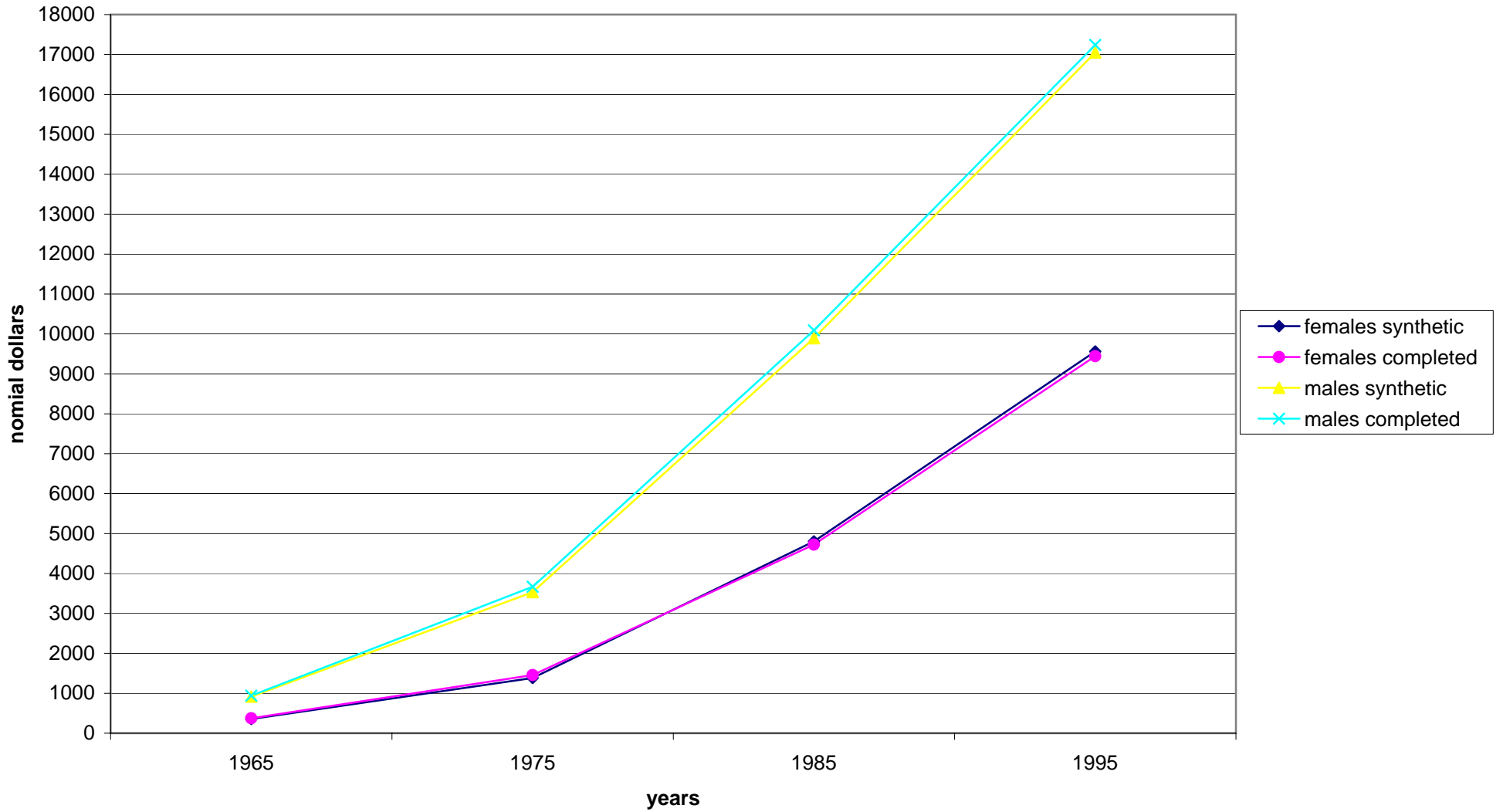


Chart 6:
Comparison of Synthetic and Completed Annual Work Indicators
Black Males and Females



**Chart 7:
Comparison of Synthetic and Completed Earnings
White Males and Females**



**Chart 8:
Comparison of Synthetic and Completed Earnings
Black Males and Females**

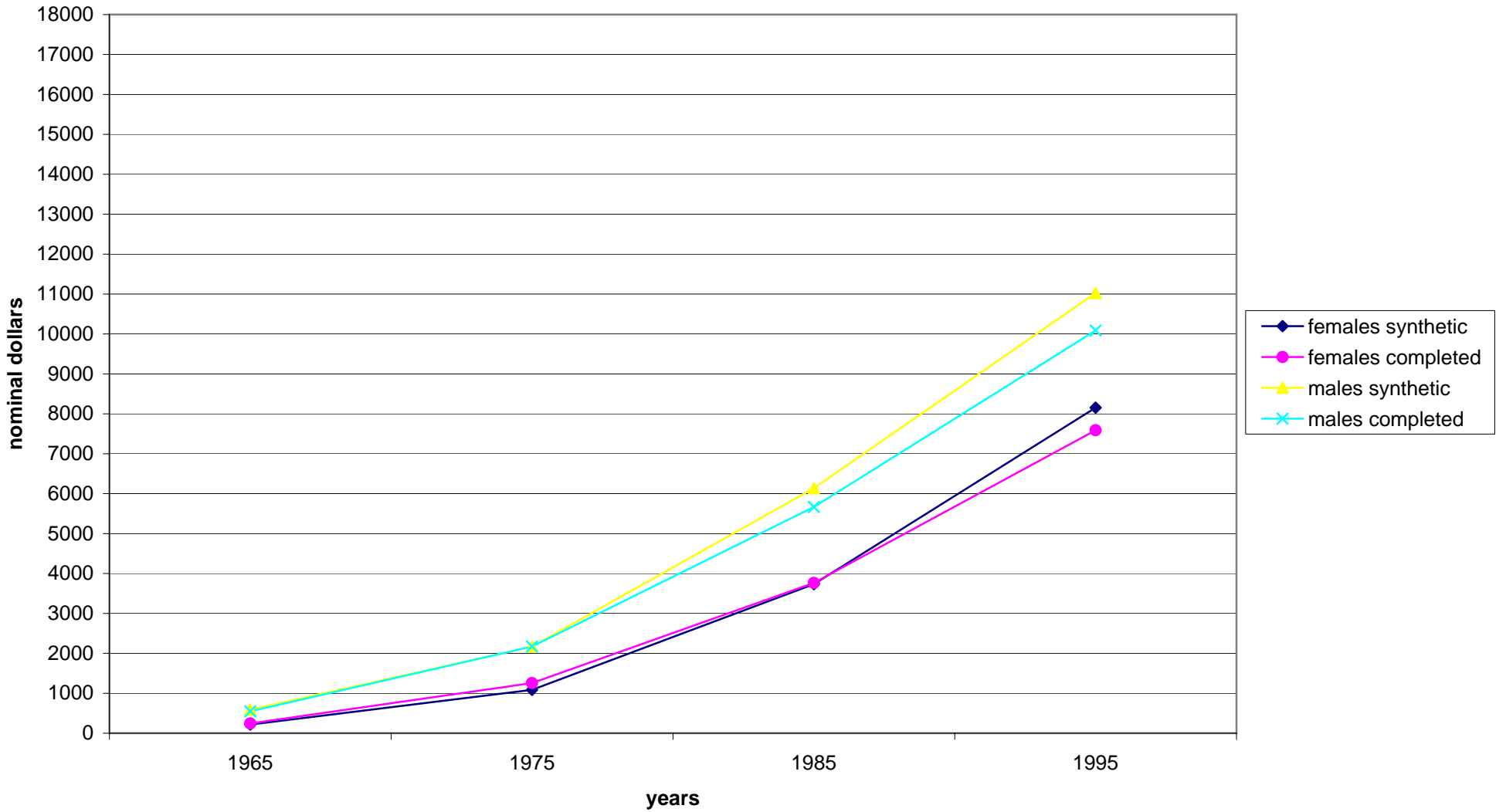


Table 29: MBA 2000 by demographic group and education

Demographic Group	Education Category	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	no HS	583	563	573	594	558	569	0	27	12
	HS	581	598	564	598	591	605	0	59	16
	Some Coll	560	598	551	569	590	606	0	24	22
	College	542	584	530	554	567	602	0	48	98
	Graduate	594	643	581	607	626	660	0	61	109
black females	no HS	485	469	472	498	457	481	0	54	53
	HS	442	445	429	456	431	459	0	55	69
	Some Coll	430	448	400	460	418	477	0	229	254
	College	444	450	414	474	408	492	0	305	634
	Graduate	507	620	415	598	558	682	0	1,891	1,313
white males	no HS	715	709	682	749	700	717	0	214	25
	HS	719	739	685	753	731	747	0	205	23
	Some Coll	708	745	682	735	734	755	0	138	40
	College	786	812	758	814	796	828	0	200	91
	Graduate	844	886	796	893	869	904	0	421	106
black males	no HS	598	581	563	633	562	599	0	229	118
	HS	538	522	489	587	501	544	0	472	167
	Some Coll	514	490	487	540	462	518	0	228	276
	College	567	584	519	616	490	678	0	837	2,682
	Graduate	701	650	574	828	583	717	0	3,801	1,652

Table 30: Non-deferred DER earnings at FICA covered jobs in year 2000

Demographic Group	Education Category	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	no HS	14,347	12,127	13,370	15,323	11,900	12,354	1	329,876	18,339
	HS	19,365	18,087	18,346	20,385	17,889	18,286	0	158,922	14,163
	Some Coll	23,788	22,750	22,951	24,626	22,478	23,021	1	242,767	26,524
	College	37,113	40,244	35,104	39,122	35,498	44,990	1	1,396,219	8,326,165
	Graduate	41,993	45,673	39,748	44,238	44,110	47,237	1	1,743,860	903,424
black females	no HS	12,228	10,116	11,594	12,862	9,716	10,515	1	139,051	58,343
	HS	17,665	15,890	15,755	19,574	15,474	16,306	1	1,261,484	63,757
	Some Coll	22,279	20,658	20,540	24,018	20,180	21,136	1	1,046,677	84,419
	College	34,678	34,163	28,613	40,743	32,760	35,566	1	12,729,526	726,972
	Graduate	42,786	46,089	24,999	60,572	39,969	52,209	0	54,819,308	13,842,902
white males	no HS	22,488	19,493	20,356	24,620	19,107	19,878	0	889,127	50,091
	HS	33,480	28,981	31,743	35,217	28,618	29,344	0	457,599	48,371
	Some Coll	41,935	37,778	36,181	47,689	35,693	39,863	1	11,455,543	1,606,545
	College	73,805	71,557	60,892	86,718	67,724	75,390	0	39,656,678	5,415,726
	Graduate	87,676	97,780	52,174	123,177	87,226	108,334	0	244,178,702	41,171,337
black males	no HS	17,240	13,829	15,180	19,299	13,111	14,547	0	795,733	185,727
	HS	24,834	23,625	22,925	26,744	21,257	25,993	1	1,261,829	2,071,917
	Some Coll	30,040	26,578	25,958	34,123	25,562	27,593	1	5,767,275	380,971
	College	51,602	46,514	33,670	69,533	41,896	51,133	0	88,347,426	7,792,256
	Graduate	79,714	64,460	-35,258	194,687	48,561	80,360	0	4,023,002,613	93,392,302

Table 31: Total SER Earnings in year 2000

Demographic Group	Education Category	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	no HS	6,475	5,421	5,379	7,572	5,299	5,543	0	197,635	5,258
	HS	10,623	10,048	10,383	10,862	9,911	10,185	1	19,877	6,592
	Some Coll	15,443	15,187	15,255	15,630	15,016	15,358	1	12,137	10,790
	College	21,319	21,206	20,733	21,904	20,834	21,579	1	118,553	50,338
	Graduate	23,213	25,799	22,519	23,907	25,311	26,286	1	166,698	87,334
black females	no HS	6,039	5,047	5,536	6,541	4,839	5,256	1	87,461	15,884
	HS	10,918	9,722	10,459	11,376	9,414	10,031	1	72,827	34,296
	Some Coll	15,324	14,827	14,667	15,980	14,427	15,227	1	149,168	59,123
	College	21,304	21,630	17,955	24,654	20,476	22,785	0	1,674,893	491,591
	Graduate	23,047	24,907	19,996	26,097	22,977	26,836	0	1,835,723	1,304,147
white males	no HS	12,534	11,598	10,828	14,241	11,373	11,824	0	488,189	17,105
	HS	20,307	18,559	19,893	20,722	18,343	18,776	1	59,358	16,462
	Some Coll	25,595	25,041	24,795	26,396	24,781	25,302	1	221,947	24,726
	College	34,103	35,830	31,265	36,942	35,385	36,275	0	1,325,716	72,845
	Graduate	33,841	38,602	32,121	35,560	38,094	39,110	0	561,532	95,213
black males	no HS	8,890	7,496	7,216	10,565	7,085	7,907	0	516,139	57,999
	HS	15,135	13,730	14,669	15,601	13,268	14,192	1	75,131	77,575
	Some Coll	18,978	18,796	18,181	19,774	18,118	19,475	1	219,702	167,113
	College	27,267	27,542	22,751	31,783	25,375	29,710	0	4,519,161	1,576,041
	Graduate	26,890	29,630	23,752	30,029	27,396	31,864	0	2,186,459	1,834,061

Table 32: Foreign Born Indicator

Demographic Group	Education Category	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed			
white females	no HS	0.200	0.235	0.195	0.205	0.229	0.241	1	0.000009	0.000011
	HS	0.092	0.096	0.088	0.096	0.094	0.099	1	0.000006	0.000002
	Some Coll	0.091	0.089	0.088	0.095	0.084	0.095	1	0.000003	0.000009
	College	0.121	0.119	0.114	0.127	0.114	0.124	0	0.000007	0.000009
	Graduate	0.105	0.102	0.100	0.110	0.096	0.108	1	0.000008	0.000013
black females	no HS	0.058	0.063	0.048	0.067	0.055	0.071	0	0.000011	0.000022
	HS	0.059	0.068	0.055	0.064	0.062	0.074	0	0.000004	0.000013
	Some Coll	0.063	0.060	0.060	0.066	0.053	0.067	1	0.000003	0.000018
	College	0.108	0.099	0.082	0.135	0.076	0.123	1	0.000245	0.000171
	Graduate	0.100	0.088	0.046	0.154	0.071	0.105	0	0.000433	0.000111
white males	no HS	0.204	0.232	0.196	0.212	0.227	0.237	0	0.000011	0.000009
	HS	0.109	0.109	0.052	0.165	0.088	0.130	0	0.000568	0.000085
	Some Coll	0.093	0.094	0.084	0.102	0.090	0.098	0	0.000012	0.000006
	College	0.111	0.104	0.107	0.115	0.097	0.112	1	0.000005	0.000017
	Graduate	0.125	0.131	0.116	0.134	0.125	0.138	0	0.000014	0.000016
black males	no HS	0.069	0.067	0.059	0.079	0.059	0.076	0	0.000024	0.000026
	HS	0.075	0.075	0.057	0.092	0.062	0.088	0	0.000059	0.000045
	Some Coll	0.078	0.079	0.070	0.086	0.069	0.089	1	0.000022	0.000034
	College	0.120	0.119	0.102	0.138	0.092	0.146	0	0.000095	0.000240
	Graduate	0.116	0.145	0.074	0.159	0.121	0.170	0	0.000320	0.000223

Table 33: Hispanic Indicator

Demographic Group	Education Category	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	no HS	0.239	0.270	0.227	0.250	0.265	0.275	0	0.000016	0.000009
	HS	0.094	0.091	0.089	0.099	0.089	0.093	0	0.000003	0.000002
	Some Coll	0.088	0.082	0.083	0.092	0.079	0.084	0	0.000003	0.000003
	College	0.050	0.040	0.047	0.053	0.037	0.043	1	0.000004	0.000003
	Graduate	0.037	0.032	0.033	0.040	0.029	0.035	1	0.000004	0.000003
black females	no HS	0.043	0.048	0.035	0.051	0.043	0.053	0	0.000016	0.000010
	HS	0.033	0.034	0.028	0.038	0.030	0.038	0	0.000008	0.000006
	Some Coll	0.041	0.036	0.030	0.053	0.031	0.040	0	0.000028	0.000008
	College	0.034	0.031	0.023	0.044	0.021	0.040	0	0.000035	0.000032
	Graduate	0.033	0.029	0.021	0.044	0.019	0.039	0	0.000046	0.000036
white males	no HS	0.264	0.288	0.246	0.282	0.283	0.293	0	0.000064	0.000009
	HS	0.116	0.114	0.109	0.122	0.111	0.117	0	0.000006	0.000003
	Some Coll	0.090	0.090	0.087	0.093	0.087	0.093	0	0.000003	0.000003
	College	0.047	0.040	0.042	0.052	0.038	0.043	0	0.000006	0.000003
	Graduate	0.040	0.036	0.035	0.045	0.033	0.039	0	0.000006	0.000003
black males	no HS	0.046	0.053	0.042	0.050	0.048	0.059	0	0.000006	0.000012
	HS	0.038	0.037	0.034	0.043	0.032	0.041	0	0.000008	0.000008
	Some Coll	0.037	0.036	0.031	0.043	0.030	0.042	0	0.000011	0.000013
	College	0.046	0.029	0.035	0.058	0.019	0.039	0	0.000046	0.000037
	Graduate	0.041	0.030	0.030	0.053	0.018	0.042	0	0.000045	0.000051

Table 34: SIPP Disability Indicator

Demographic Group	Education Category	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white females	no HS	0.179	0.186	0.173	0.185	0.181	0.192	0	0.000008	0.000011
	HS	0.121	0.108	0.117	0.125	0.104	0.112	1	0.000006	0.000006
	Some Coll	0.088	0.081	0.080	0.095	0.077	0.085	1	0.000020	0.000005
	College	0.062	0.050	0.032	0.091	0.046	0.053	0	0.000157	0.000004
	Graduate	0.064	0.053	0.061	0.067	0.049	0.057	1	0.000003	0.000006
black females	no HS	0.199	0.223	0.183	0.216	0.203	0.242	0	0.000083	0.000115
	HS	0.125	0.133	0.114	0.136	0.122	0.143	0	0.000026	0.000037
	Some Coll	0.087	0.086	0.078	0.097	0.079	0.094	0	0.000026	0.000021
	College	0.052	0.032	0.030	0.073	0.021	0.042	0	0.000115	0.000039
	Graduate	0.099	0.057	0.065	0.133	0.041	0.073	0	0.000276	0.000096
white males	no HS	0.164	0.164	0.162	0.167	0.154	0.174	1	0.000003	0.000026
	HS	0.112	0.112	0.111	0.114	0.107	0.116	1	0.000001	0.000006
	Some Coll	0.089	0.085	0.081	0.097	0.081	0.089	0	0.000008	0.000006
	College	0.051	0.043	0.048	0.055	0.039	0.047	0	0.000004	0.000005
	Graduate	0.061	0.048	0.047	0.075	0.044	0.052	0	0.000033	0.000006
black males	no HS	0.189	0.200	0.175	0.204	0.186	0.213	0	0.000032	0.000065
	HS	0.134	0.146	0.116	0.153	0.135	0.157	0	0.000051	0.000041
	Some Coll	0.105	0.095	0.090	0.120	0.083	0.107	0	0.000057	0.000052
	College	0.077	0.066	0.044	0.110	0.046	0.087	0	0.000242	0.000139
	Graduate	0.148	0.097	0.054	0.241	0.075	0.119	0	0.001727	0.000180

Table 35: Non-deferred DER Earnings at FICA covered jobs in year 2000

Demographic Group	Foreign Born Category	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white female	yes	11,823	10,828	9,709	13,937	10,513	11,142	0	960,850	36,442
	no	13,999	14,141	13,736	14,262	13,714	14,568	1	23,915	67,516
black female	yes	14,232	12,525	12,726	15,738	11,406	13,644	1	784,465	429,817
	no	12,360	11,128	11,283	13,436	10,836	11,419	1	401,114	31,016
white male	yes	25,102	24,808	19,252	30,952	22,599	27,016	0	7,862,844	1,798,716
	no	30,453	29,986	28,122	32,784	28,914	31,059	1	1,880,323	424,824
black male	yes	19,786	17,812	15,493	24,079	13,952	21,673	0	4,399,306	5,016,775
	no	17,553	15,275	11,840	23,266	14,446	16,104	0	9,117,175	252,176

Table 36: Total SER Earnings in year 2000

Demographic Group	Foreign Born Category	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed			
white female	yes	10,883	10,531	10,520	11,245	10,277	10,786	1	45,494	23,893
	no	13,442	13,311	13,248	13,637	13,192	13,430	1	13,081	4,876
black female	yes	13,372	12,688	13,095	13,650	11,701	13,676	1	26,624	350,092
	no	11,806	11,021	11,380	12,233	10,784	11,258	1	62,905	20,049
white male	yes	19,774	19,568	17,461	22,086	19,133	20,003	0	931,942	65,411
	no	23,352	23,494	23,219	23,486	23,335	23,654	1	6,169	9,105
black male	yes	16,930	15,775	15,848	18,011	14,076	17,475	0	415,547	903,569
	no	15,127	14,192	14,652	15,601	13,868	14,516	1	77,832	38,238

Table 37: Monthly Benefit Amount April 2000

Demographic Group	Foreign Born Category	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed	Synthetic	Completed	Synthetic	Completed		Synthetic	Completed
white female	yes	528	536	517	539	515	557	0	33	117
	no	580	595	574	587	591	599	0	12	5
black female	yes	434	494	409	460	458	529	0	232	472
	no	460	458	451	468	449	468	0	26	30
white male	yes	628	687	592	664	647	728	0	295	389
	no	747	760	742	752	755	765	0	9	10
black male	yes	490	563	448	532	499	628	1	623	1,378
	no	573	545	562	583	533	557	0	34	54

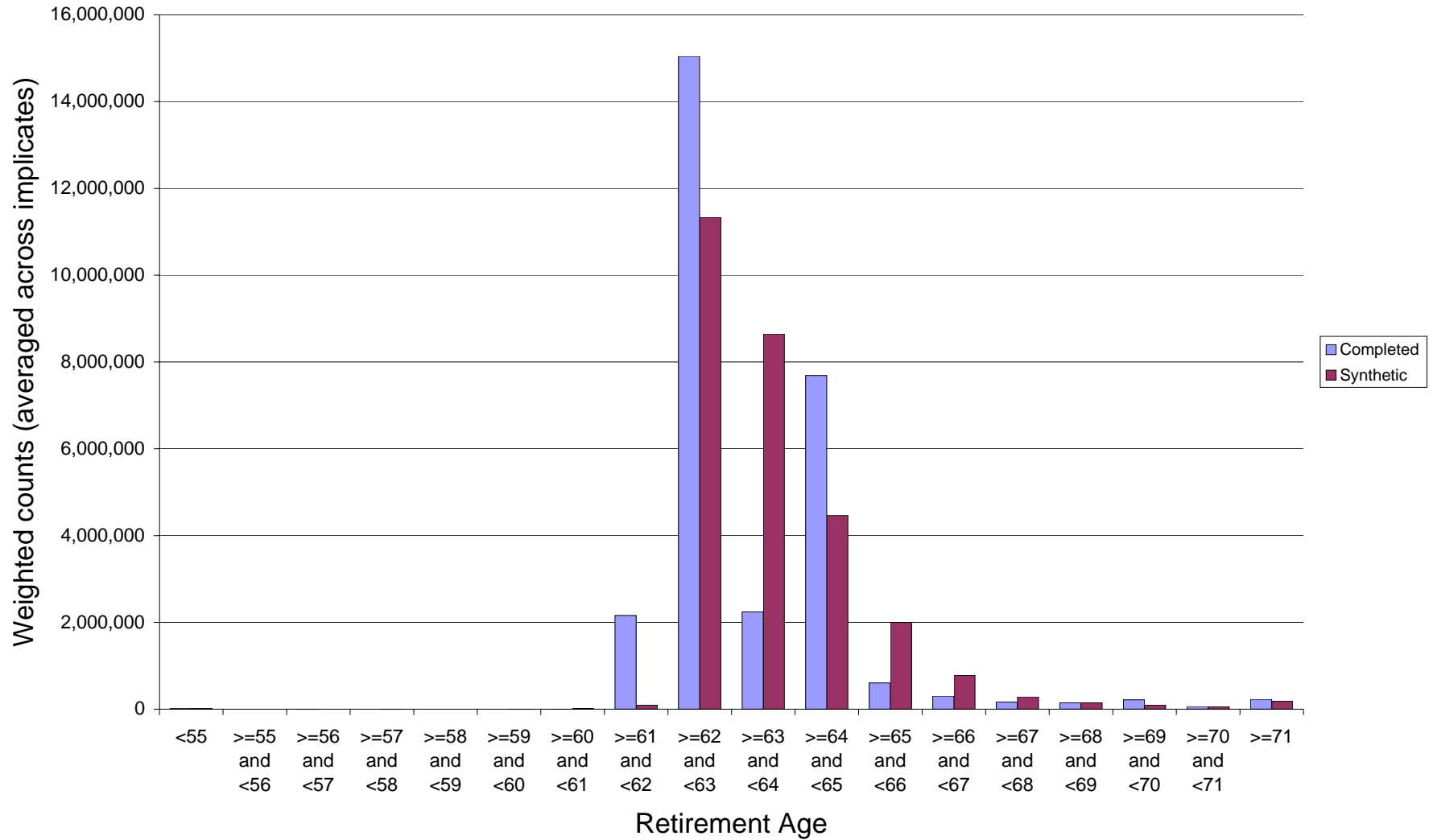
Table 38: Marital History Variables

Variable	Mean		Confidence Interval		Confidence Interval		Synthetic DF Not Exist	Total Variance	
	Synthetic	Completed	Synthetic		Completed			Synthetic	Completed
Number of marriages	0.88	0.87			0.87	0.87	1	-5.38E-08	9.68E-07
Percent ever divorced	0.24	0.24			0.24	0.24	1	1.15E-07	3.89E-07
Percent ever widowed	0.07	0.07	0.07	0.07	0.07	0.07	0	2.53E-07	1.03E-07
Duration of 1st marriage	23.17	26.18	22.50	23.83	26.07	26.29	0	9.28E-03	2.21E-03
Duration of 2nd marriage	12.96	14.41	11.31	14.61	14.14	14.68	0	0.17	0.01
Age at first marriage	23.05	23.28			23.24	23.32	1	-3.94E-03	3.13E-04

Table 39: Retirement Age – Weighted and Unweighted Counts

Age	WEIGHTED		UNWEIGHTED	
	Completed	Synthetic	Completed	Synthetic
<55	25,115	24,089	35	33
>=55 and <56	0	0	0	0
>=56 and <57	1,724	141	1	1
>=57 and <58	0	2,396	1	2
>=58 and <59	268	0	1	1
>=59 and <60	733	3,467	1	5
>=60 and <61	4,686	13,601	11	21
>=61 and <62	2,160,505	92,222	4,919	139
>=62 and <63	15,042,278	11,330,115	21,771	18,046
>=63 and <64	2,243,418	8,648,017	3,820	14,081
>=64 and <65	7,688,675	4,461,798	12,248	7,274
>=65 and <66	605,260	1,998,145	988	3,335
>=66 and <67	302,360	781,017	474	1,388
>=67 and <68	165,783	284,813	305	472
>=68 and <69	148,499	149,679	218	229
>=69 and <70	216,884	89,701	298	137
>=70 and <71	62,259	58,498	106	91
>=71	228,409	182,367	348	289

Age at Retirement, Weighted



Age at Retirement, Unweighted

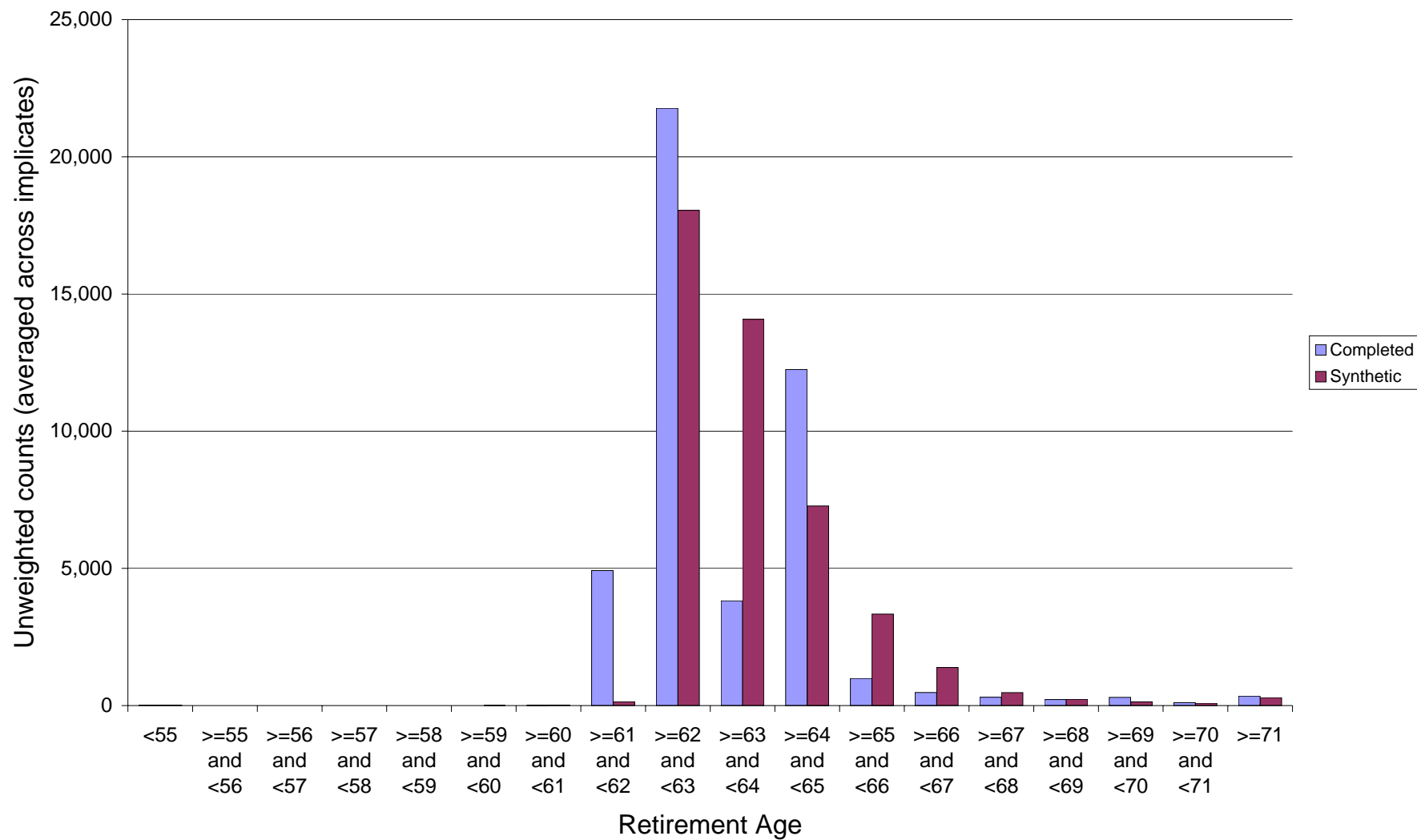


Table 40: Log of Total DER Earnings in year 2000 for white males

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	8.377	7.855	8.266	8.487	7.793	7.917	0.065	0.037	1
highschool_only	0.214	0.230	0.133	0.294	0.205	0.255	0.036	0.015	0
somecollege	0.400	0.431	0.263	0.537	0.404	0.457	0.059	0.016	0
college_only	0.738	0.880	0.530	0.947	0.851	0.909	0.086	0.017	0
graduate	0.830	1.110	0.632	1.028	1.080	1.140	0.085	0.018	0
disab	-0.354	-0.610	-0.380	-0.328	-0.657	-0.562	0.014	0.026	0
foreign_born	0.064	0.042	-0.029	0.157	0.013	0.070	0.042	0.017	0
hispanic	-0.072	-0.013	-0.113	-0.031	-0.040	0.013	0.021	0.016	0
ser_totyrs_2000	0.179	0.275	0.142	0.216	0.259	0.292	0.014	0.010	0
ser_totyrs_2000_2	-0.073	-0.140	-0.085	-0.062	-0.153	-0.128	0.007	0.007	1
ser_totyrs_2000_3	0.016	0.034	0.013	0.018	0.030	0.038	0.001	0.002	1
ser_totyrs_2000_4	-0.001	-0.003	-0.002	-0.001	-0.004	-0.003	0.000	0.000	1

Table 41: Log of Total DER Earnings in year 2000 for black males

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	8.080	7.070	7.929	8.230	6.885	7.254	0.089	0.108	1
highschool_only	0.163	0.322	-0.031	0.357	0.231	0.413	0.090	0.053	0
somecollege	0.375	0.551	0.204	0.546	0.476	0.627	0.074	0.046	0
college_only	0.680	0.860	0.415	0.945	0.735	0.985	0.124	0.075	0
graduate	0.797	1.169	0.461	1.133	1.018	1.320	0.156	0.091	0
disab	-0.400	-0.631	-0.533	-0.267	-0.763	-0.499	0.062	0.075	0
foreign_born	0.082	0.046	-0.098	0.262	-0.106	0.197	0.084	0.084	0
hispanic	-0.030	0.156	-0.128	0.067	0.017	0.296	0.051	0.084	0
ser_totyrs_2000	0.173	0.388	0.154	0.191	0.336	0.440	0.011	0.030	1
ser_totyrs_2000_2	-0.067	-0.240	-0.078	-0.055	-0.284	-0.197	0.007	0.025	1
ser_totyrs_2000_3	0.013	0.067	0.009	0.018	0.053	0.080	0.003	0.008	1
ser_totyrs_2000_4	-0.001	-0.007	-0.002	-0.001	-0.008	-0.005	0.000	0.001	1

Table 42: Log of Total DER Earnings in year 2000 for white males

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	8.373	7.717	8.295	8.451	7.658	7.777	0.046	0.036	1
highschool_only	0.180	0.248	0.157	0.203	0.221	0.274	0.013	0.016	1
somecollege	0.405	0.491	0.284	0.527	0.463	0.520	0.051	0.017	0
college_only	0.719	0.834	0.561	0.876	0.802	0.865	0.063	0.019	0
graduate	0.790	1.043	0.628	0.952	1.008	1.078	0.064	0.021	0
disab	-0.259	-0.470	-0.296	-0.222	-0.511	-0.428	0.022	0.024	1
foreign_born	0.075	0.097	0.032	0.117	0.062	0.132	0.025	0.020	1
hispanic	-0.008	0.043	-0.049	0.033	0.007	0.079	0.022	0.021	0
ser_totyrs_2000	0.104	0.216	0.039	0.169	0.201	0.230	0.025	0.009	0
ser_totyrs_2000_2	-0.038	-0.119	-0.091	0.014	-0.131	-0.107	0.020	0.007	0
ser_totyrs_2000_3	0.009	0.032	-0.007	0.025	0.029	0.036	0.006	0.002	0
ser_totyrs_2000_4	-0.001	-0.003	-0.003	0.001	-0.004	-0.003	0.001	0.000	0

Table 43: Log of Total DER Earnings in year 2000 for white males

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	8.236	7.194	8.101	8.372	7.059	7.329	0.080	0.082	1
highschool_only	0.263	0.347	0.197	0.330	0.282	0.412	0.032	0.039	0
somecollege	0.494	0.587	0.376	0.612	0.518	0.656	0.046	0.041	0
college_only	0.842	1.118	0.770	0.914	1.030	1.207	0.042	0.054	1
graduate	0.953	1.289	0.748	1.157	1.179	1.399	0.077	0.067	0
disab	-0.300	-0.453	-0.408	-0.192	-0.543	-0.363	0.052	0.054	0
foreign_born	0.114	0.183	-0.016	0.245	0.050	0.317	0.066	0.073	0
hispanic	-0.011	-0.007	-0.124	0.101	-0.113	0.099	0.064	0.064	0
ser_totyrs_2000	0.086	0.297	-0.010	0.183	0.260	0.335	0.037	0.023	0
ser_totyrs_2000_2	-0.019	-0.175	-0.092	0.054	-0.207	-0.144	0.031	0.019	0
ser_totyrs_2000_3	0.003	0.048	-0.019	0.024	0.038	0.059	0.010	0.006	0
ser_totyrs_2000_4	0.000	-0.005	-0.003	0.002	-0.006	-0.004	0.001	0.001	0

Table 44: Log of SER Earnings in year 2000 for white males

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	8.277	7.807	8.135	8.419	7.746	7.867	0.084	0.036	1
highschool_only	0.190	0.198	0.102	0.278	0.175	0.221	0.038	0.014	0
somecollege	0.357	0.366	0.246	0.467	0.342	0.390	0.048	0.015	0
college_only	0.596	0.700	0.430	0.762	0.672	0.727	0.071	0.016	0
graduate	0.638	0.794	0.458	0.819	0.765	0.823	0.077	0.018	0
disab	-0.340	-0.580	-0.364	-0.316	-0.634	-0.527	0.014	0.028	0
foreign_born	0.015	0.042	-0.086	0.116	0.017	0.067	0.045	0.015	0
hispanic	-0.041	-0.028	-0.090	0.007	-0.053	-0.003	0.024	0.015	0
ser_totyrs_2000	0.194	0.282	0.150	0.238	0.266	0.298	0.016	0.009	0
ser_totyrs_2000_2	-0.087	-0.148	-0.102	-0.071	-0.160	-0.136	0.009	0.007	1
ser_totyrs_2000_3	0.019	0.036	0.016	0.023	0.033	0.040	0.002	0.002	1
ser_totyrs_2000_4	-0.002	-0.003	-0.002	-0.001	-0.004	-0.003	0.000	0.000	1
married	0.111	0.110	0.078	0.145	0.088	0.132	0.016	0.013	0
divorced	-0.072	-0.042	-0.094	-0.050	-0.073	-0.010	0.012	0.019	0
widowed	-0.357	-0.423	-0.599	-0.115	-0.545	-0.302	0.120	0.074	0

Table 45: Log of SER earnings in year 2000 for black males

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	8.022	7.106	7.861	8.183	6.931	7.281	0.095	0.105	1
highschool_only	0.130	0.292	-0.049	0.308	0.194	0.390	0.084	0.056	0
somecollege	0.317	0.500	0.186	0.449	0.421	0.579	0.058	0.048	0
college_only	0.557	0.658	0.329	0.785	0.525	0.792	0.109	0.080	0
graduate	0.598	0.872	0.333	0.864	0.717	1.027	0.129	0.094	0
disab	-0.367	-0.571	-0.515	-0.218	-0.702	-0.439	0.067	0.076	0
foreign_born	0.032	0.066	-0.140	0.204	-0.096	0.228	0.079	0.090	0
hispanic	0.001	0.187	-0.107	0.108	0.050	0.325	0.056	0.083	0
ser_totyrs_2000	0.172	0.359	0.131	0.213	0.311	0.407	0.024	0.029	1
ser_totyrs_2000_2	-0.066	-0.215	-0.098	-0.034	-0.255	-0.175	0.019	0.024	1
ser_totyrs_2000_3	0.013	0.059	0.003	0.023	0.047	0.072	0.006	0.008	1
ser_totyrs_2000_4	-0.001	-0.006	-0.002	0.000	-0.007	-0.005	0.001	0.001	1
married	0.114	0.097	0.057	0.171	0.022	0.172	0.034	0.045	0
divorced	-0.187	-0.190	-0.278	-0.095	-0.340	-0.040	0.053	0.085	0
widowed	-0.311	-0.492	-0.716	0.093	-0.999	0.016	0.223	0.293	0

Table 46: Log of SER Earnings in year 2000 for white females

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	8.300	7.641	8.216	8.385	7.579	7.702	0.050	0.037	1
highschool_only	0.168	0.242	0.134	0.202	0.215	0.269	0.020	0.016	1
somecollege	0.374	0.431	0.244	0.503	0.403	0.460	0.055	0.017	0
college_only	0.646	0.710	0.471	0.820	0.677	0.742	0.071	0.020	0
graduate	0.686	0.833	0.531	0.840	0.797	0.869	0.059	0.022	0
disab	-0.249	-0.445	-0.302	-0.195	-0.486	-0.403	0.019	0.024	0
foreign_born	0.076	0.194	0.006	0.147	0.159	0.229	0.025	0.020	0
hispanic	0.019	0.030	-0.027	0.066	-0.010	0.070	0.024	0.023	0
ser_totyrs_2000	0.111	0.229	0.043	0.179	0.214	0.244	0.027	0.009	0
ser_totyrs_2000_2	-0.028	-0.109	-0.081	0.025	-0.121	-0.096	0.021	0.008	0
ser_totyrs_2000_3	0.004	0.027	-0.011	0.020	0.023	0.031	0.006	0.002	0
ser_totyrs_2000_4	-0.001	-0.003	-0.002	0.001	-0.003	-0.002	0.001	0.000	0
married	-0.241	-0.342	-0.275	-0.208	-0.370	-0.314	0.020	0.016	1
divorced	-0.190	-0.243	-0.265	-0.116	-0.276	-0.210	0.029	0.020	0
widowed	-0.473	-0.563	-0.571	-0.376	-0.636	-0.491	0.049	0.042	0

Table 47: Log of SER Earnings in year 2000 for black females

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	8.107	7.150	7.970	8.244	7.006	7.294	0.081	0.087	1
highschool_only	0.230	0.307	0.152	0.307	0.244	0.370	0.036	0.038	0
somecollege	0.439	0.493	0.325	0.553	0.424	0.562	0.047	0.041	0
college_only	0.733	0.913	0.625	0.842	0.818	1.007	0.064	0.057	1
graduate	0.811	1.013	0.619	1.002	0.899	1.126	0.085	0.069	0
disab	-0.284	-0.407	-0.375	-0.193	-0.501	-0.312	0.047	0.056	0
foreign_born	0.094	0.229	-0.053	0.241	0.102	0.357	0.072	0.071	0
hispanic	0.008	0.033	-0.109	0.125	-0.074	0.139	0.067	0.065	0
ser_totyrs_2000	0.102	0.292	0.041	0.162	0.252	0.333	0.036	0.024	1
ser_totyrs_2000_2	-0.021	-0.159	-0.098	0.056	-0.193	-0.126	0.029	0.020	0
ser_totyrs_2000_3	0.003	0.043	-0.019	0.024	0.032	0.054	0.009	0.007	0
ser_totyrs_2000_4	0.000	-0.004	-0.002	0.002	-0.006	-0.003	0.001	0.001	0
married	-0.073	-0.091	-0.126	-0.021	-0.157	-0.025	0.031	0.039	0
divorced	-0.170	-0.214	-0.218	-0.123	-0.294	-0.133	0.028	0.047	0
widowed	-0.435	-0.469	-0.606	-0.264	-0.601	-0.336	0.094	0.081	0

Table 48: Log of Average Indexed Monthly Earnings (AIME) or Average Monthly Wage (AMW) for all individuals

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic		Completed		Synthetic	Completed	DF Not Exist
Intercept	7.604	7.252	7.554	7.654	7.170	7.335	0.029	0.048	1
age_2000	0.0004	0.0093	-0.0067	0.0075	0.0055	0.0131	0.003	0.002	0
age_2000_sq	-0.0002	-0.0003	-0.0003	-0.0001	-0.0003	-0.0002	0.000	0.000	0
blackfemale	-0.928	-0.995	-0.949	-0.906	-1.019	-0.972	0.010	0.014	0
blackmale	-0.403	-0.457	-0.444	-0.362	-0.499	-0.415	0.019	0.022	0
whitefemale	-0.822	-0.843	-0.836	-0.807	-0.853	-0.832	0.007	0.006	0
highschool_only	0.337	0.400	0.235	0.438	0.382	0.417	0.043	0.010	0
somecollege	0.570	0.690	0.441	0.699	0.673	0.708	0.055	0.010	0
college_only	0.717	0.866	0.571	0.862	0.840	0.891	0.062	0.014	0
graduate	0.748	0.911	0.641	0.855	0.879	0.942	0.046	0.017	0
disab	-0.365	-0.559	-0.488	-0.241	-0.580	-0.538	0.053	0.012	0
hispanic	-0.249	-0.257	-0.280	-0.218	-0.276	-0.237	0.014	0.011	0
divorced	0.136	0.159	0.108	0.164	0.118	0.200	0.015	0.021	0
married	0.134	0.132	0.105	0.162	0.099	0.165	0.014	0.017	0
widowed	-0.106	-0.024	-0.145	-0.067	-0.062	0.014	0.022	0.023	0

*AIME for individuals who turned 62 after 1979, AMW otherwise

Table 49: Log of initial MBA for retired individuals (TOB_initial=1)

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	-61.534	-67.501	-66.344	-56.725	-71.471	-63.531	2.501	2.192	0
age_initial_entitle	0.033	0.038	0.028	0.038	0.033	0.044	0.003	0.003	0
blackfemale	-0.360	-0.329	-0.435	-0.285	-0.386	-0.272	0.036	0.030	0
blackmale	-0.110	-0.120	-0.150	-0.071	-0.150	-0.089	0.015	0.018	0
whitefemale	-0.301	-0.297	-0.364	-0.238	-0.354	-0.240	0.027	0.026	0
highschool_only	0.070	0.061	0.042	0.097	0.026	0.096	0.014	0.017	0
somecollege	0.121	0.089	0.078	0.163	0.054	0.124	0.020	0.018	0
college_only	0.164	0.124	0.143	0.184	0.080	0.168	0.011	0.022	0
graduate	0.191	0.147	0.150	0.232	0.119	0.175	0.020	0.016	0
disab	-0.048	-0.039	-0.076	-0.021	-0.067	-0.011	0.013	0.015	0
hispanic	-0.098	-0.058	-0.161	-0.035	-0.124	0.009	0.029	0.032	0
divorced	0.114	0.132	0.069	0.159	0.098	0.166	0.023	0.020	0
married	0.078	0.052	0.052	0.104	0.019	0.085	0.015	0.019	0
widowed	0.179	0.162	0.146	0.213	0.126	0.197	0.020	0.021	0
log_totnetworth	0.015	0.046	0.005	0.025	0.037	0.055	0.005	0.005	0
ser_pct_yrs_wrked	1.052	1.044	0.609	1.496	0.689	1.398	0.187	0.151	0
year_initial_entitle	0.033	0.035	0.030	0.035	0.033	0.037	0.001	0.001	0

Table 50: Log of initial MBA for disabled individuals (TOB_initial=2)

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	-76.179	-75.378	-80.608	-71.751	-80.502	-70.253	2.283	2.663	0
age_initial_entitle	0.010	0.010	0.009	0.011	0.009	0.012	0.001	0.001	0
blackfemale	-0.299	-0.255	-0.353	-0.244	-0.321	-0.188	0.029	0.035	0
blackmale	-0.055	-0.022	-0.093	-0.018	-0.059	0.015	0.021	0.022	0
whitefemale	-0.328	-0.326	-0.347	-0.309	-0.348	-0.304	0.011	0.013	0
highschool_only	0.070	0.143	0.028	0.113	0.106	0.180	0.021	0.020	0
somecollege	0.139	0.201	0.102	0.176	0.168	0.233	0.020	0.019	0
college_only	0.213	0.287	0.164	0.262	0.225	0.349	0.023	0.034	0
graduate	0.233	0.343	0.192	0.274	0.287	0.399	0.023	0.033	0
disab	-0.045	-0.004	-0.068	-0.022	-0.025	0.017	0.012	0.013	0
hispanic	-0.058	-0.047	-0.091	-0.025	-0.096	0.001	0.019	0.027	0
divorced	0.099	0.101	0.067	0.131	0.070	0.133	0.019	0.019	0
married	0.125	0.133	0.102	0.149	0.097	0.170	0.014	0.021	0
widowed	0.046	0.067	0.003	0.088	0.018	0.116	0.025	0.030	0
log_totnetworth	0.005	0.016	-0.002	0.011	0.010	0.022	0.003	0.004	0
ser_pct_yrs_wrked	0.542	0.536	0.351	0.734	0.388	0.683	0.085	0.069	0
year_initial_entitle	0.041	0.040	0.039	0.043	0.038	0.043	0.001	0.001	0

Table 51: Log of MBA 2000 for retired individuals (TOB_2000=1)

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	21.365	16.985	11.656	31.074	7.475	26.496	4.487	4.459	0
age_2000	0.009	0.009	0.005	0.012	0.007	0.011	0.002	0.001	0
blackfemale	-0.272	-0.250	-0.339	-0.206	-0.301	-0.199	0.032	0.026	0
blackmale	-0.095	-0.092	-0.122	-0.068	-0.127	-0.056	0.014	0.020	0
whitefemale	-0.211	-0.192	-0.284	-0.139	-0.257	-0.126	0.032	0.029	0
highschool_only	0.054	0.054	0.039	0.068	0.024	0.084	0.007	0.015	0
somecollege	0.098	0.078	0.075	0.121	0.052	0.104	0.012	0.014	0
college_only	0.136	0.128	0.103	0.169	0.084	0.172	0.016	0.022	0
graduate	0.157	0.150	0.132	0.182	0.125	0.175	0.013	0.015	0
disab	-0.019	0.001	-0.035	-0.003	-0.019	0.021	0.009	0.011	0
hispanic	-0.102	-0.059	-0.164	-0.040	-0.108	-0.009	0.028	0.025	0
divorced	0.110	0.126	0.071	0.148	0.075	0.177	0.021	0.027	0
married	0.107	0.081	0.073	0.142	0.051	0.110	0.019	0.017	0
widowed	0.232	0.217	0.187	0.278	0.171	0.264	0.024	0.026	0
famwelpart1999	-0.048	-0.022	-0.103	0.008	-0.058	0.013	0.026	0.019	0
hicoannual1999	0.023	-0.001	0.013	0.033	-0.012	0.010	0.006	0.007	0
log_totnetworth	0.012	0.040	0.004	0.021	0.035	0.046	0.004	0.003	0
ser_pct_yrs_wrked	0.948	1.001	0.473	1.423	0.593	1.409	0.202	0.174	0
year_initial_entitle	-0.008	-0.006	-0.013	-0.003	-0.011	-0.001	0.002	0.002	0

Table 52: Log of MBA 2000 for disabled individuals (TOB_2000=2)

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	-15.944	-18.763	-19.131	-12.758	-22.303	-15.222	1.849	2.129	0
age_2000	0.010	0.010	0.009	0.011	0.009	0.012	0.001	0.001	0
blackfemale	-0.260	-0.226	-0.319	-0.201	-0.284	-0.168	0.033	0.033	0
blackmale	-0.048	-0.008	-0.086	-0.011	-0.056	0.041	0.022	0.028	0
whitefemale	-0.293	-0.295	-0.314	-0.271	-0.326	-0.264	0.013	0.018	0
highschool_only	0.065	0.111	0.024	0.106	0.073	0.149	0.019	0.022	0
somecollege	0.126	0.175	0.093	0.158	0.141	0.209	0.018	0.020	0
college_only	0.189	0.280	0.152	0.226	0.204	0.357	0.021	0.043	0
graduate	0.207	0.280	0.161	0.253	0.215	0.345	0.026	0.039	0
disab	-0.040	0.019	-0.072	-0.008	-0.007	0.045	0.016	0.015	0
hispanic	-0.034	-0.057	-0.073	0.005	-0.097	-0.017	0.023	0.024	0
divorced	0.037	0.029	-0.025	0.099	-0.020	0.077	0.031	0.028	0
married	0.067	0.065	0.029	0.104	0.014	0.115	0.021	0.028	0
widowed	-0.009	0.028	-0.076	0.059	-0.048	0.103	0.039	0.045	0
famwelpart1999	-0.032	-0.017	-0.072	0.008	-0.059	0.025	0.022	0.024	0
hicoannual1999	0.038	0.017	0.007	0.069	-0.013	0.046	0.017	0.018	0
log_totnetworth	0.002	0.019	-0.006	0.010	0.012	0.026	0.004	0.004	0
ser_pct_yrs_wrked	0.413	0.379	0.318	0.507	0.288	0.470	0.048	0.049	0
year_initial_entitle	0.011	0.012	0.009	0.013	0.010	0.014	0.001	0.001	0

Table 53: Log of Total family income in year 1999, all individuals

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	-0.491	3.914	-1.852	0.870	-0.420	8.249	0.801	1.905	1
blackfemale	-0.354	-0.450	-0.418	-0.291	-0.545	-0.356	0.038	0.041	1
blackmale	-0.239	-0.293	-0.349	-0.129	-0.620	0.034	0.065	0.139	1
whitefemale	-0.156	-0.171	-0.254	-0.058	-0.504	0.161	0.058	0.142	1
highschool_only	0.118	0.211	-0.040	0.275	0.059	0.364	0.055	0.065	0
somecollege	0.252	0.402	0.164	0.340	0.282	0.522	0.052	0.051	1
college_only	0.354	0.715	0.321	0.387	0.446	0.983	0.020	0.114	1
graduate	0.412	0.895	0.375	0.449	0.490	1.300	0.022	0.172	1
disab	-0.121	-0.252	-0.151	-0.091	-0.288	-0.215	0.013	0.017	0
foreign_born	-0.051	-0.060	-0.092	-0.011	-0.098	-0.022	0.018	0.017	0
hispanic	-0.117	-0.150	-0.186	-0.048	-0.220	-0.081	0.030	0.031	0
divorced	-0.190	-0.176	-0.248	-0.132	-0.800	0.447	0.034	0.265	1
married	0.485	0.420	0.407	0.564	0.206	0.634	0.046	0.091	1
widowed	-0.031	-0.035	-0.091	0.030	-0.443	0.374	0.036	0.174	1
retired	-0.236	-0.141	-0.316	-0.156	-0.171	-0.111	0.029	0.015	0
disabledssa	-0.353	-0.202	-0.412	-0.295	-0.245	-0.158	0.025	0.022	0
agedspouse	-0.291	-0.248	-0.351	-0.230	-0.434	-0.061	0.036	0.083	1
widowspouse	-0.305	-0.244	-0.376	-0.234	-0.349	-0.139	0.028	0.050	0
otherbenefit	-0.111	-0.092	-0.145	-0.077	-0.131	-0.053	0.016	0.018	0
totfam_kids	-0.054	0.027	-0.057	-0.051	0.009	0.045	0.002	0.008	1
year_birth	0.006	0.003	0.005	0.006	0.001	0.005	0.000	0.001	1

Table 54: Log of total personal income in year 1999 for all individuals

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	40.724	39.397	36.384	45.064	31.231	47.564	2.553	3.529	1
blackfemale	-0.663	-0.574	-0.873	-0.454	-0.716	-0.433	0.073	0.061	0
blackmale	-0.381	-0.326	-0.411	-0.350	-0.354	-0.299	0.018	0.016	1
whitefemale	-0.705	-0.693	-0.913	-0.497	-1.021	-0.365	0.086	0.140	0
highschool_only	0.247	0.299	0.140	0.354	0.260	0.338	0.042	0.018	0
somecollege	0.423	0.452	0.302	0.543	0.387	0.518	0.048	0.029	0
college_only	0.722	0.922	0.502	0.941	0.684	1.160	0.087	0.102	0
graduate	0.787	1.165	0.537	1.036	0.835	1.495	0.097	0.141	0
disab	-0.155	-0.313	-0.188	-0.121	-0.356	-0.270	0.015	0.020	0
foreign_born	-0.063	-0.101	-0.116	-0.010	-0.125	-0.077	0.021	0.013	0
hispanic	-0.149	-0.058	-0.229	-0.069	-0.087	-0.029	0.033	0.015	0
divorced	0.254	0.223	-0.018	0.527	0.039	0.407	0.096	0.079	0
married	0.246	0.211	0.214	0.278	0.095	0.326	0.019	0.050	1
widowed	0.233	0.335	0.187	0.278	0.142	0.529	0.027	0.085	1
retired	-0.622	-0.527	-0.742	-0.502	-0.639	-0.415	0.051	0.049	0
disabledssa	-0.405	-0.273	-0.470	-0.340	-0.351	-0.194	0.029	0.037	0
agedspouse	-1.011	-1.017	-1.187	-0.835	-1.173	-0.862	0.074	0.073	0
widowspouse	-0.658	-0.634	-0.830	-0.487	-0.882	-0.386	0.076	0.110	0
otherbenefit	-0.054	-0.005	-0.095	-0.013	-0.052	0.043	0.019	0.023	0
totfam_kids	0.004	-0.037	-0.020	0.028	-0.053	-0.021	0.009	0.007	0
year_birth	-0.016	-0.015	-0.018	-0.014	-0.019	-0.011	0.001	0.002	1

Table 55: Indicator for whether individual has either a DB or DC pension, all individuals age/employment eligible f

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	-4.573	-4.999	-4.633	-4.514	-5.227	-4.771	0.035	0.097	1
age_2000	0.049	0.079	0.043	0.055	0.065	0.093	0.004	0.006	1
age_2000_sq	-0.0003	-0.0006	-0.0004	-0.0002	-0.0008	-0.0005	0.000	0.000	1
blackfemale	-0.112	-0.125	-0.281	0.057	-0.319	0.070	0.064	0.083	0
blackmale	0.077	0.046	0.071	0.083	-0.068	0.161	0.003	0.049	1
whitefemale	-0.213	-0.246	-0.284	-0.142	-0.314	-0.178	0.028	0.029	0
highschool_only	0.304	0.387	0.217	0.391	0.376	0.398	0.031	0.005	0
somecollege	0.492	0.579	0.368	0.617	0.528	0.629	0.048	0.022	0
college_only	0.744	0.804	0.612	0.877	0.742	0.865	0.050	0.026	0
graduate	0.656	0.678	0.525	0.788	0.666	0.690	0.047	0.005	0
disab	-0.063	-0.205	-0.123	-0.003	-0.308	-0.102	0.035	0.044	1
hispanic	-0.187	-0.261	-0.291	-0.083	-0.324	-0.199	0.041	0.027	0
divorced	0.084	0.121	0.047	0.120	0.063	0.178	0.021	0.024	1
married	0.198	0.257	0.178	0.218	0.209	0.305	0.012	0.020	1
widowed	-0.063	0.016	-0.135	0.009	-0.203	0.234	0.042	0.093	1
ltotearn_ser_2000	0.254	0.229	0.244	0.264	0.208	0.251	0.006	0.009	1
managerial	0.177	0.315	0.087	0.268	0.295	0.335	0.032	0.009	0
tech_support	0.085	0.208	0.039	0.132	0.191	0.226	0.027	0.007	1
manufacturing	0.144	0.292	0.135	0.153	0.231	0.352	0.005	0.026	1
retail	0.011	-0.349	-0.025	0.047	-0.407	-0.291	0.021	0.025	1
services	0.041	-0.063	0.016	0.066	-0.123	-0.003	0.015	0.026	1

Table 56: Indicator for positive weeks with pay in year 1999, all individuals

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	0.227	0.061	0.186	0.268	-0.255	0.378	0.024	0.134	1
blackfemale	-0.404	-0.393	-0.547	-0.262	-0.504	-0.281	0.052	0.047	0
blackmale	-0.389	-0.486	-0.548	-0.230	-0.591	-0.382	0.059	0.045	0
whitefemale	-0.346	-0.387	-0.360	-0.332	-0.426	-0.348	0.008	0.017	1
highschool_only	0.571	0.795	0.547	0.595	0.630	0.960	0.014	0.070	1
somecollege	0.775	0.921	0.749	0.800	0.788	1.054	0.015	0.057	1
college_only	0.961	1.245	0.902	1.019	1.150	1.340	0.035	0.040	1
graduate	0.919	1.214	0.860	0.978	1.123	1.305	0.035	0.039	1
totfam_kids	0.161	0.155	0.123	0.199	0.127	0.182	0.015	0.012	0
foreign_born	0.008	0.095	-0.127	0.143	-0.017	0.206	0.056	0.047	0
hispanic	0.146	0.176	0.115	0.176	0.090	0.262	0.018	0.037	1
pos_mba2000	-1.148	-1.525	-1.167	-1.128	-1.601	-1.448	0.011	0.033	1
der_avg_log_real_ear	0.024	-0.011	0.020	0.029	-0.020	-0.003	0.003	0.004	1
qder2	0.199	0.292	0.162	0.236	0.194	0.391	0.022	0.042	1
qder3	0.323	0.439	0.201	0.445	0.341	0.536	0.046	0.041	0
qder4	0.257	0.278	0.087	0.427	0.151	0.405	0.069	0.054	0
ser_pct_yrs_wrked	0.307	1.179	-0.273	0.888	0.188	2.170	0.238	0.421	0
qser2	0.182	0.110	0.012	0.352	-0.209	0.429	0.070	0.135	0
qser3	0.704	0.567	0.419	0.988	0.060	1.074	0.117	0.215	0
qser4	1.102	1.105	0.739	1.465	0.521	1.689	0.150	0.248	0
divorced	-0.264	-0.289	-0.295	-0.232	-0.395	-0.184	0.018	0.045	1
divorced1	0.086	0.128	0.075	0.098	0.100	0.157	0.007	0.012	1
married	-0.498	-0.522	-0.574	-0.423	-0.594	-0.451	0.029	0.030	0
widowed	-0.984	-0.955	-1.028	-0.940	-1.059	-0.852	0.026	0.044	1
widowed1	-0.623	-0.472	-0.667	-0.579	-0.530	-0.413	0.026	0.025	1

Table 57: Log of total networth in 2000 dollars, one observation per married couple

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	67.110	75.883	60.560	73.660	68.672	83.094	2.928	3.352	0
black	-0.654	-0.315	-0.827	-0.481	-0.474	-0.155	0.070	0.090	0
black_spouse	-0.026	-0.294	-0.164	0.111	-0.446	-0.142	0.064	0.087	0
highschool_only	0.482	0.315	0.464	0.500	0.283	0.346	0.011	0.019	1
somecollege	0.634	0.443	0.565	0.703	0.399	0.486	0.040	0.025	1
college_only	1.057	0.769	0.959	1.155	0.718	0.819	0.058	0.029	1
graduate	1.152	0.841	1.054	1.250	0.788	0.894	0.058	0.030	1
highschool_only_spouse	0.031	0.290	0.002	0.060	0.258	0.323	0.018	0.019	0
somecollege_spouse	0.042	0.398	-0.003	0.087	0.366	0.431	0.025	0.020	0
college_only_spouse	0.074	0.677	0.040	0.107	0.641	0.713	0.019	0.022	0
graduate_spouse	0.083	0.696	0.051	0.115	0.652	0.740	0.019	0.026	0
totfam_kids	-0.039	-0.042	-0.062	-0.016	-0.067	-0.017	0.010	0.012	0
foreign_born	0.036	-0.001	0.007	0.066	-0.051	0.048	0.018	0.028	0
foreign_born_spouse	0.012	0.061	-0.018	0.042	-0.010	0.133	0.018	0.037	0
hispanic	-0.068	-0.168	-0.119	-0.017	-0.263	-0.074	0.029	0.051	0
hispanic_spouse	-0.056	-0.126	-0.092	-0.020	-0.217	-0.035	0.018	0.049	0
own_home	1.349	1.765	1.326	1.371	1.712	1.818	0.013	0.027	1
retired	0.024	-0.089	-0.033	0.081	-0.138	-0.039	0.024	0.028	0
disabledssa	-0.163	-0.253	-0.348	0.022	-0.317	-0.188	0.085	0.038	0
agedspouse	0.023	-0.089	-0.061	0.106	-0.174	-0.003	0.049	0.050	1
widowspouse	-0.034	-0.282	-0.220	0.151	-0.389	-0.174	0.073	0.060	0
otherbenefit	-0.011	-0.058	-0.052	0.030	-0.100	-0.015	0.024	0.025	0
retired_spouse	0.032	0.079	-0.023	0.087	0.038	0.120	0.028	0.024	0
disabledssa_spouse	-0.026	-0.079	-0.122	0.070	-0.135	-0.023	0.047	0.034	0
agedspouse_spouse	0.045	0.104	-0.042	0.132	0.033	0.175	0.047	0.043	0
widowspouse_spouse	-0.061	-0.024	-0.206	0.083	-0.200	0.151	0.070	0.095	0
otherbenefit_spouse	-0.019	-0.039	-0.064	0.026	-0.087	0.010	0.025	0.028	0
year_birth	-0.026	-0.021	-0.028	-0.024	-0.026	-0.016	0.001	0.002	1
year_birth_spouse	-0.003	-0.014	-0.007	0.000	-0.016	-0.011	0.002	0.001	1

Table 58: Log of total net worth in 2000 dollars, single individuals

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	16.892	15.110	14.374	19.410	11.606	18.613	1.481	1.773	1
male	0.073	0.067	0.015	0.132	0.000	0.133	0.034	0.030	1
black	-0.687	-0.646	-0.775	-0.599	-0.798	-0.493	0.052	0.067	1
highschool_only	0.057	0.148	-0.021	0.136	0.074	0.222	0.046	0.034	1
somecollege	0.198	0.392	0.110	0.285	0.320	0.465	0.051	0.034	1
college_only	0.376	0.645	0.316	0.436	0.586	0.704	0.035	0.031	1
graduate	0.460	0.778	0.393	0.526	0.690	0.866	0.039	0.044	1
widowed	-0.070	-0.079	-0.107	-0.032	-0.138	-0.019	0.022	0.033	1
divorced	-0.301	-0.346	-0.353	-0.250	-0.429	-0.263	0.030	0.038	1
foreign_born	0.032	0.036	-0.013	0.077	-0.050	0.121	0.026	0.041	1
totfam_kids	0.009	-0.017	-0.004	0.021	-0.027	-0.007	0.007	0.006	1
hispanic	-0.132	-0.367	-0.187	-0.076	-0.429	-0.305	0.024	0.031	0
own_home	1.774	2.126	1.745	1.803	2.090	2.163	0.017	0.018	1
retired	0.232	0.312	0.184	0.279	0.266	0.357	0.028	0.027	1
disabledssa	-0.069	-0.159	-0.133	-0.005	-0.221	-0.097	0.035	0.036	0
agedspouse	0.137	0.309	-0.155	0.428	0.033	0.586	0.151	0.150	0
otherbenefit	-0.043	-0.110	-0.093	0.006	-0.155	-0.065	0.025	0.024	0
widowspouse	0.157	0.289	0.066	0.247	0.235	0.343	0.035	0.033	0
year_birth	-0.004	-0.003	-0.005	-0.003	-0.005	-0.001	0.001	0.001	1

Table 59: Log of home equity in 2000 dollars, one observation per married couple

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	57.204	68.769	52.663	61.746	62.126	75.411	2.671	3.172	1
black	-0.271	-0.127	-0.673	0.130	-0.257	0.002	0.148	0.078	0
black_spouse	-0.028	-0.269	-0.147	0.090	-0.395	-0.144	0.055	0.076	0
highschool_only	0.218	0.134	0.064	0.373	0.067	0.202	0.091	0.034	1
somecollege	0.314	0.187	0.156	0.471	0.123	0.251	0.093	0.034	1
college_only	0.552	0.335	0.449	0.654	0.215	0.455	0.061	0.057	1
graduate	0.577	0.383	0.460	0.695	0.313	0.454	0.069	0.038	1
highschool_only_spouse	0.012	0.140	-0.016	0.041	0.104	0.177	0.016	0.021	0
somecollege_spouse	0.010	0.188	-0.032	0.053	0.150	0.226	0.020	0.023	0
college_only_spouse	0.021	0.309	-0.034	0.076	0.262	0.355	0.028	0.027	0
graduate_spouse	0.032	0.313	-0.021	0.086	0.255	0.371	0.027	0.033	0
foreign_born	0.126	0.090	0.071	0.182	0.019	0.161	0.029	0.037	0
foreign_born_spouse	0.000	0.069	-0.028	0.028	0.014	0.125	0.015	0.031	0
totfam_kids	-0.015	-0.001	-0.029	-0.002	-0.025	0.023	0.007	0.012	0
hispanic	-0.087	-0.066	-0.128	-0.047	-0.133	0.001	0.023	0.040	0
hispanic_spouse	-0.001	-0.121	-0.043	0.040	-0.184	-0.058	0.024	0.038	0
qder2	0.016	-0.004	-0.016	0.049	-0.043	0.036	0.019	0.023	0
qder3	-0.008	0.016	-0.054	0.038	-0.022	0.053	0.024	0.022	0
qder4	0.060	0.178	-0.005	0.125	0.133	0.223	0.039	0.025	1
qder2_spouse	0.013	0.016	-0.029	0.055	-0.013	0.046	0.023	0.018	0
qder3_spouse	0.031	0.023	-0.016	0.078	-0.007	0.053	0.023	0.018	0
qder4_spouse	0.091	0.219	0.040	0.143	0.188	0.250	0.030	0.019	1
qser2	-0.002	-0.036	-0.048	0.043	-0.076	0.004	0.024	0.022	0
qser3	-0.027	-0.077	-0.077	0.024	-0.133	-0.021	0.026	0.029	0
qser4	-0.062	-0.103	-0.153	0.029	-0.180	-0.026	0.039	0.038	0
qser2_spouse	-0.008	-0.028	-0.054	0.037	-0.063	0.007	0.016	0.020	0
qser3_spouse	0.004	-0.088	-0.037	0.045	-0.126	-0.050	0.021	0.022	0
qser4_spouse	-0.001	-0.120	-0.094	0.093	-0.175	-0.065	0.034	0.029	0
retired	0.012	-0.083	-0.029	0.054	-0.120	-0.047	0.017	0.022	0
disabledssa	-0.043	-0.063	-0.277	0.190	-0.135	0.010	0.104	0.042	0
agedspouse	0.008	-0.052	-0.066	0.082	-0.128	0.023	0.043	0.045	0
widowspouse	-0.031	-0.237	-0.116	0.055	-0.338	-0.136	0.045	0.057	0
otherbenefit	-0.002	-0.028	-0.044	0.039	-0.078	0.021	0.023	0.029	0
retired_spouse	0.034	0.046	0.015	0.054	0.005	0.088	0.012	0.024	1
disabledssa_spouse	0.030	0.008	-0.024	0.084	-0.052	0.068	0.032	0.036	0
agedspouse_spouse	0.044	0.125	-0.045	0.133	0.044	0.205	0.048	0.047	0
widowspouse_spouse	0.012	-0.026	-0.073	0.097	-0.157	0.104	0.035	0.075	0
otherbenefit_spouse	-0.033	-0.020	-0.079	0.014	-0.081	0.041	0.028	0.034	0
year_birth	-0.021	-0.019	-0.024	-0.019	-0.021	-0.018	0.001	0.001	1
year_birth_spouse	-0.003	-0.011	-0.006	0.001	-0.014	-0.007	0.002	0.002	1

Table 60: Log of home equity in 2000 dollars, single individuals

Explanatory Variables	Coefficient		Confidence Interval				Standard Error		Synthetic
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	DF Not Exist
Intercept	23.169	14.041	16.960	29.378	9.200	18.881	3.652	2.360	1
male	0.042	0.019	-0.112	0.196	-0.152	0.190	0.091	0.074	1
black	-0.296	-0.437	-0.507	-0.085	-0.600	-0.274	0.124	0.072	1
highschool_only	0.047	0.103	-0.004	0.099	0.065	0.142	0.031	0.021	1
somecollege	0.201	0.244	0.054	0.348	0.219	0.269	0.086	0.015	1
college_only	0.301	0.354	0.220	0.382	0.310	0.397	0.048	0.025	1
graduate	0.328	0.407	0.243	0.412	0.351	0.462	0.050	0.032	1
foreign_born	0.069	0.066	-0.019	0.157	0.009	0.124	0.039	0.031	0
hispanic	-0.111	-0.222	-0.166	-0.055	-0.316	-0.129	0.026	0.045	0
totfam_kids	-0.023	-0.012	-0.038	-0.007	-0.024	0.000	0.009	0.007	1
qder2	0.020	-0.002	-0.039	0.079	-0.037	0.034	0.023	0.020	0
qder3	0.014	-0.060	-0.009	0.037	-0.100	-0.020	0.014	0.022	1
qder4	0.023	0.045	-0.008	0.055	-0.024	0.113	0.015	0.035	0
qser2	-0.011	-0.023	-0.058	0.036	-0.054	0.008	0.018	0.018	0
qser3	-0.022	-0.072	-0.089	0.046	-0.116	-0.029	0.031	0.023	0
qser4	-0.024	-0.099	-0.102	0.054	-0.148	-0.049	0.033	0.026	0
retired	0.119	0.271	0.101	0.138	0.212	0.330	0.011	0.033	1
disabledssa	-0.005	-0.016	-0.064	0.053	-0.076	0.045	0.028	0.036	0
agedspouse	0.120	0.218	-0.099	0.340	0.007	0.429	0.113	0.123	0
widowspouse	0.130	0.274	0.109	0.152	0.200	0.348	0.012	0.043	1
otherbenefit	-0.017	-0.047	-0.073	0.039	-0.093	-0.001	0.020	0.025	0
divorced	-0.326	-0.324	-0.349	-0.304	-0.487	-0.161	0.013	0.072	1
widowed	-0.053	-0.024	-0.160	0.054	-0.198	0.149	0.063	0.079	1
year_birth	-0.006	-0.002	-0.009	-0.003	-0.004	0.001	0.002	0.001	1

Table 61: Percentiles of Synthetic and Completed Variables

Panel 1

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
Date variables											
birthdate	completed	1/22/1955	1/12/1913	4/28/1922	9/6/1928	4/21/1943	6/13/1957	4/1/1969	2/1/1977	9/10/1979	4/20/1981
birthdate	synthesized	2/17/1955	4/24/1913	8/22/1922	3/23/1929	10/1/1943	7/2/1957	1/27/1969	8/25/1976	6/10/1979	3/7/1981
date_initial_entitle	completed	3/9/1988	12/9/1963	1/31/1970	10/9/1973	12/24/1980	10/24/1989	5/24/1996	6/1/2000	9/9/2001	9/1/2002
date_initial_entitle	synthesized	5/17/1988	3/3/1964	4/5/1970	11/21/1973	3/7/1981	12/21/1989	7/30/1996	6/20/2000	8/31/2001	9/29/2002
deathdate	completed	7/5/2001	4/12/2000	5/17/2000	7/16/2000	12/2/2000	7/3/2001	2/17/2002	6/24/2002	8/5/2002	9/14/2002
deathdate	synthesized	10/19/2000	2/4/1993	4/22/1996	8/6/1998	7/13/2000	3/18/2001	11/26/2001	6/2/2002	8/28/2002	12/7/2002
MBA Variables											
mba_2000	completed	643	28	91	156	353	611	919	1136	1260	1549
mba_2000	synthesized	642	37	94	155	350	609	921	1137	1259	1537
mba_initial_real	completed	609	34	105	165	337	549	873	1116	1236	1433
mba_initial_real	synthesized	612	45	110	170	339	551	880	1120	1237	1431
Marital History Variables											
age_mar1	completed	23.4	15.8	17.2	18.1	19.8	22.3	25.6	30	33.4	42.4
age_mar1	synthesized	23.1	15.7	17	17.9	19.6	22.1	25.2	29.2	32.4	40.9
duration_end1	completed	928	0	1	1	3	13	1961	1973	1977	1981
duration_end1	synthesized	963	0	1	2	4	242	1948	1975	2005	2070
duration_end2	completed	1074	0	1	2	4	1933	1960	1968	1972	1978
duration_end2	synthesized	1059	0	2	4	21	1860	1951	1988	2015	2062
duration_end3	completed	1814	1	6	1928	1942	1953	1961	1967	1969	1974
duration_end3	synthesized	1842	1	723	1929	1944	1954	1961	1966	1969	1973
duration_mar1	completed	14.5	0.3	1.2	2.2	4.7	9.6	20.1	36	44.7	55.5
duration_mar1	synthesized	13.4	0.3	1.3	2.2	4.4	8.8	18.2	33.1	42.3	53.6
duration_mar2	completed	1169	0	2	4	9	1955	1970	1975	1978	1981
duration_mar2	synthesized	1200	1	2	5	17	1941	1969	1976	1983	2065
duration_mar3	completed	1298	0	2	3	10	1953	1964	1970	1973	1979
duration_mar3	synthesized	1218	2	11	27	141	1803	1965	1987	2026	2113
duration_mar4	completed	1955	1925	1936	1941	1949	1956	1962	1967	1970	1972
duration_mar4	synthesized	1956	1924	1934	1942	1950	1957	1963	1968	1970	1973
Wealth Variables											
homeequity	completed	72314	-9000	4000	8000	22000	50000	100000	163625	215750	320000
homeequity	synthesized	74491	-26272	1327	5284	18539	48942	101062	178872	249326	380942
nonhouswealth	completed	74925	-7000	1000	2000	6000	17000	61000	181500	317250	765000
nonhouswealth	synthesized	72921	-75235	-513	1056	4695	15181	56042	177601	324071	877694
totnetworth	completed	119632	-33000	-6000	1000	9000	51500	141000	294500	449500	925000
totnetworth	synthesized	113145	-42844	-7631	-995	7525	49761	137418	287070	436222	879640
DER Earnings Arrays											
nondefer_der_fica_1978	completed	13200	49	311	738	2540	7188	14026	21072	27729	66966
nondefer_der_fica_1978	synthesized	14111	111	516	996	2691	7106	13879	21139	28327	83151
nondefer_der_fica_1979	completed	13268	56	345	813	2745	7730	14933	22900	28896	58653
nondefer_der_fica_1979	synthesized	14352	69	394	943	2898	7726	15020	22903	28958	63122
nondefer_der_fica_1980	completed	11930	58	349	825	2872	8272	16014	25068	30336	53290
nondefer_der_fica_1980	synthesized	12041	171	729	1338	3369	8306	15972	24643	30185	56505
nondefer_der_fica_1981	completed	12601	61	397	969	3299	9315	17711	27405	33374	56929
nondefer_der_fica_1981	synthesized	12565	109	539	1157	3481	9218	17576	27084	32833	56298
nondefer_der_fica_1982	completed	13439	70	439	1020	3586	10071	18752	28941	35401	60673
nondefer_der_fica_1982	synthesized	13426	170	731	1415	3767	10007	18400	28326	34782	60943

Table 61: Percentiles of Synthetic and Completed Variables

Panel 2

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
nondefer_der_fica_1983	completed	14184	68	445	1046	3687	10434	19686	30530	37532	65243
nondefer_der_fica_1983	synthesized	14091	121	622	1348	3774	10201	19032	29762	36987	64334
nondefer_der_fica_1984	completed	15053	77	471	1104	3975	11105	20862	32730	39905	69921
nondefer_der_fica_1984	synthesized	15393	180	818	1571	4142	11262	20622	32266	39665	75406
nondefer_der_fica_1985	completed	15815	70	479	1141	4044	11514	21920	34376	41800	74477
nondefer_der_fica_1985	synthesized	15998	137	658	1431	4190	11504	21660	33907	42649	78072
nondefer_der_fica_1986	completed	16667	78	492	1154	4221	12019	22930	36112	44478	81420
nondefer_der_fica_1986	synthesized	17291	140	685	1467	4302	11864	22824	35629	44217	87617
nondefer_der_fica_1987	completed	17392	79	500	1181	4385	12538	23720	37322	45232	84994
nondefer_der_fica_1987	synthesized	17681	157	785	1608	4603	12255	23284	37089	46097	89174
nondefer_der_fica_1988	completed	18259	85	536	1253	4570	13066	24759	39211	47747	89655
nondefer_der_fica_1988	synthesized	18254	88	547	1408	4589	12705	24179	38614	47761	90015
nondefer_der_fica_1989	completed	18877	94	572	1340	4862	13641	25744	40601	49613	93885
nondefer_der_fica_1989	synthesized	18771	146	727	1563	4747	13386	25392	40108	49767	93357
nondefer_der_fica_1990	completed	19588	90	596	1429	5122	14180	26505	41143	51300	96543
nondefer_der_fica_1990	synthesized	19555	156	814	1784	5139	13968	26231	40710	52098	98197
nondefer_der_fica_1991	completed	20495	94	599	1443	5295	14548	27211	42615	55136	109567
nondefer_der_fica_1991	synthesized	20754	130	681	1614	5244	14354	26864	42141	55638	114055
nondefer_der_fica_1992	completed	21543	91	581	1410	5356	15033	28443	44740	58130	120450
nondefer_der_fica_1992	synthesized	21869	143	735	1724	5277	14807	28218	44329	58774	121441
nondefer_der_fica_1993	completed	22267	91	578	1459	5595	15449	29289	46379	61093	127753
nondefer_der_fica_1993	synthesized	22631	214	995	2079	5892	15331	28937	46081	61796	127778
nondefer_der_fica_1994	completed	22634	88	572	1455	5597	15633	29561	46731	61417	125867
nondefer_der_fica_1994	synthesized	22989	215	980	2067	5931	15428	29325	46729	62059	130595
nondefer_der_fica_1995	completed	23562	90	600	1488	5771	16082	30322	48258	63701	131903
nondefer_der_fica_1995	synthesized	23918	276	1220	2420	6580	16215	30160	48015	64170	135003
nondefer_der_fica_1996	completed	25237	97	670	1587	6000	16655	31301	49953	66417	138301
nondefer_der_fica_1996	synthesized	25106	201	985	2209	6658	17160	31587	50411	67309	146184
nondefer_der_fica_1997	completed	25871	106	784	1856	6513	17645	32966	52836	70687	152864
nondefer_der_fica_1997	synthesized	26333	180	922	2196	6828	17719	33212	53691	72785	153970
nondefer_der_fica_1998	completed	27732	133	924	2170	7365	18928	34784	55510	74645	159782
nondefer_der_fica_1998	synthesized	28554	272	1260	2704	7840	19285	35359	55974	75283	163610
nondefer_der_fica_1999	completed	29647	140	1058	2553	8273	20198	36406	58082	78723	170779
nondefer_der_fica_1999	synthesized	31594	340	1648	3409	9017	20806	36978	58745	79890	181358
nondefer_der_fica_2000	completed	32320	151	1249	2970	9484	21767	38395	61340	83931	182894
nondefer_der_fica_2000	synthesized	33828	390	1930	3957	10297	22575	39376	63049	87879	189390
nondefer_der_fica_2001	completed	33095	156	1263	3204	10368	23159	40185	63882	86996	184584
nondefer_der_fica_2001	synthesized	34710	390	1961	4209	10922	23799	40856	64637	88257	197413
nondefer_der_fica_2002	completed	33697	123	1204	3245	10645	24004	41408	65771	88989	188232
nondefer_der_fica_2002	synthesized	35702	343	1753	4047	11018	24395	42237	67194	91413	195640
nondefer_der_fica_2003	completed	34757	132	1323	3364	11208	24997	42835	67701	91212	193449
nondefer_der_fica_2003	synthesized	37130	451	2194	4633	11531	24988	43179	68410	94524	224561
nondefer_der_nonfica_1978	completed	8535	19	75	185	817	5845	13604	20211	25742	38178
nondefer_der_nonfica_1978	synthesized	8650	46	207	394	1149	5700	13533	19893	24992	37818
nondefer_der_nonfica_1979	completed	8845	17	81	198	869	6411	14545	21144	25320	36151
nondefer_der_nonfica_1979	synthesized	9100	101	422	758	1830	6303	14472	20851	25309	36236
nondefer_der_nonfica_1980	completed	9901	21	100	248	1119	7384	16079	23074	28059	40143

Table 61: Percentiles of Synthetic and Completed Variables

Panel 3

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
nondefer_der_nonfica_1980	synthesized	10160	122	470	817	1919	6769	16321	23521	28542	41489
nondefer_der_nonfica_1981	completed	10949	22	86	211	871	6844	17912	26556	32848	50107
nondefer_der_nonfica_1981	synthesized	10773	106	462	818	1910	6326	17168	25717	31693	48010
nondefer_der_nonfica_1982	completed	11272	16	71	165	644	4563	19983	28826	35213	55306
nondefer_der_nonfica_1982	synthesized	11236	93	386	679	1594	4743	19015	28530	35113	53224
nondefer_der_nonfica_1983	completed	12299	19	73	188	765	6945	21869	30197	36284	53715
nondefer_der_nonfica_1983	synthesized	12079	107	459	825	2029	6653	20131	28886	35534	51606
nondefer_der_nonfica_1984	completed	13853	18	75	169	759	9137	24213	32116	38660	57567
nondefer_der_nonfica_1984	synthesized	13061	104	415	742	1775	5959	22783	31979	38904	56810
nondefer_der_nonfica_1985	completed	13502	19	72	163	717	6852	24628	33471	39735	57649
nondefer_der_nonfica_1985	synthesized	15486	89	371	679	1890	10668	26453	35924	43316	60674
nondefer_der_nonfica_1986	completed	14329	20	75	188	797	7225	25513	34544	41974	66503
nondefer_der_nonfica_1986	synthesized	14020	90	374	686	1677	5848	24703	34201	41875	65683
nondefer_der_nonfica_1987	completed	14429	23	83	195	813	7237	25952	35488	42813	62477
nondefer_der_nonfica_1987	synthesized	14067	112	460	830	2047	6390	24155	35012	42592	60991
nondefer_der_nonfica_1988	completed	15509	21	87	204	842	8517	27618	37607	45095	65608
nondefer_der_nonfica_1988	synthesized	15077	92	336	561	1231	6690	26713	37364	45505	68137
nondefer_der_nonfica_1989	completed	16660	20	83	209	838	8282	28640	39336	47914	68857
nondefer_der_nonfica_1989	synthesized	15346	82	352	635	1464	6077	27323	39244	48224	70655
nondefer_der_nonfica_1990	completed	16199	25	92	218	902	7892	29256	40921	49527	70694
nondefer_der_nonfica_1990	synthesized	15421	104	400	692	1514	5460	27815	40621	49230	70348
nondefer_der_nonfica_1991	completed	18196	26	99	245	1080	11200	31943	44379	53115	74783
nondefer_der_nonfica_1991	synthesized	17499	104	460	837	2268	8152	30653	43700	52676	71958
nondefer_der_nonfica_1992	completed	19096	24	104	257	1123	11862	33221	46387	55533	80737
nondefer_der_nonfica_1992	synthesized	17644	121	491	884	2187	7697	30607	44997	54274	76585
nondefer_der_nonfica_1993	completed	20172	30	123	314	1293	11553	34179	48740	59780	92424
nondefer_der_nonfica_1993	synthesized	19030	125	506	900	2140	8289	32277	47649	59339	92763
nondefer_der_nonfica_1994	completed	20348	25	111	300	1267	11755	34253	48882	58945	87604
nondefer_der_nonfica_1994	synthesized	18970	133	573	1087	2817	9103	32023	47427	56973	80100
nondefer_der_nonfica_1995	completed	20954	28	113	285	1306	11744	34788	50428	62353	100654
nondefer_der_nonfica_1995	synthesized	19603	130	565	1013	2304	8460	32854	48237	59353	98481
nondefer_der_nonfica_1996	completed	21164	33	112	266	1205	11268	35392	51058	62305	95121
nondefer_der_nonfica_1996	synthesized	19590	142	592	1076	2620	8226	32561	49626	61014	86993
nondefer_der_nonfica_1997	completed	22308	35	144	341	1508	12205	36578	53045	66063	107457
nondefer_der_nonfica_1997	synthesized	22394	181	753	1398	3444	10221	34013	53722	67143	107630
nondefer_der_nonfica_1998	completed	22258	47	173	379	1485	11700	36887	53559	66271	107811
nondefer_der_nonfica_1998	synthesized	21579	196	797	1419	3317	10101	34584	53356	66338	103551
nondefer_der_nonfica_1999	completed	23507	40	177	435	1761	12787	37658	55287	67788	107295
nondefer_der_nonfica_1999	synthesized	23865	250	1031	1855	4418	12217	36606	55908	70316	110034
nondefer_der_nonfica_2000	completed	24279	43	177	441	1779	12922	39062	57009	69929	114863
nondefer_der_nonfica_2000	synthesized	23886	273	1030	1813	4088	11996	37009	57250	71600	112647
nondefer_der_nonfica_2001	completed	25025	35	168	454	1906	13745	40160	58644	72237	111589
nondefer_der_nonfica_2001	synthesized	25101	290	1164	2088	4939	13479	38351	59250	73265	111949
nondefer_der_nonfica_2002	completed	28247	46	227	600	2858	19476	43875	63539	78023	123295
nondefer_der_nonfica_2002	synthesized	27569	164	764	1500	4578	18092	42560	62350	77421	114900
nondefer_der_nonfica_2003	completed	29124	45	228	649	3371	21096	45665	65283	78806	124983
nondefer_der_nonfica_2003	synthesized	33294	481	2072	3793	9930	28197	48682	67110	82069	117449

Table 61: Percentiles of Synthetic and Completed Variables

Panel 4

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
defer_der_fica_1987	completed	1105	410	410	410	870	1101	1226	1637	2048	2048
defer_der_fica_1987	synthesized	923	615	615	653	775	921	1070	1132	1132	1243
defer_der_fica_1988	completed	7771	0	122	141	590	1761	3672	5874	7627	7627
defer_der_fica_1988	synthesized	13545	8	60	155	585	1942	4217	6256	77453	289657
defer_der_fica_1989	completed	3263	10	89	157	530	1200	2454	4751	6178	26488
defer_der_fica_1989	synthesized	2577	35	135	239	555	1298	2720	4993	7479	20550
defer_der_fica_1990	completed	2113	35	148	261	600	1327	2800	4997	6665	8400
defer_der_fica_1990	synthesized	1965	30	131	247	560	1188	2548	4844	6514	8803
defer_der_fica_1991	completed	2321	44	154	266	610	1370	2821	5127	7024	8750
defer_der_fica_1991	synthesized	2119	44	180	307	608	1283	2657	5065	6914	9688
defer_der_fica_1992	completed	2280	46	173	290	644	1443	2999	5429	7313	8728
defer_der_fica_1992	synthesized	2275	69	229	361	674	1388	2914	5458	7535	9529
defer_der_fica_1993	completed	2448	45	172	300	678	1503	3190	5763	7774	9491
defer_der_fica_1993	synthesized	2365	61	219	361	696	1458	3091	5753	7916	9693
defer_der_fica_1994	completed	2704	44	181	312	701	1584	3334	6003	8051	9240
defer_der_fica_1994	synthesized	2535	51	218	370	723	1574	3224	5853	7938	9584
defer_der_fica_1995	completed	2775	46	171	312	708	1619	3454	6279	8341	9240
defer_der_fica_1995	synthesized	2532	55	245	424	848	1561	3334	6113	8312	9542
defer_der_fica_1996	completed	2641	50	194	336	765	1714	3634	6672	8719	9500
defer_der_fica_1996	synthesized	2639	59	243	424	839	1725	3574	6574	8652	9711
defer_der_fica_1997	completed	2796	49	191	338	786	1812	3892	7185	9369	9500
defer_der_fica_1997	synthesized	2790	60	251	436	867	1818	3856	7075	9157	9627
defer_der_fica_1998	completed	2979	51	205	368	849	1927	4199	7722	9936	10000
defer_der_fica_1998	synthesized	2986	68	284	483	941	1941	4112	7548	9557	10591
defer_der_fica_1999	completed	3140	48	203	378	899	2024	4453	8096	10000	10000
defer_der_fica_1999	synthesized	3178	59	243	439	925	2017	4378	7975	9796	12646
defer_der_fica_2000	completed	3429	45	205	378	905	2085	4646	8574	10500	10500
defer_der_fica_2000	synthesized	3443	70	297	507	1024	2161	4688	8542	10326	11741
defer_der_fica_2001	completed	3374	45	201	387	939	2154	4824	9033	10500	10500
defer_der_fica_2001	synthesized	3429	92	338	547	1073	2225	4875	8925	10351	11178
defer_der_fica_2002	completed	3489	45	200	390	953	2197	4920	9644	11000	12000
defer_der_fica_2002	synthesized	3513	83	310	515	1037	2194	4856	9489	10973	12522
defer_der_fica_2003	completed	3720	55	225	408	1000	2271	5119	10452	12000	14000
defer_der_fica_2003	synthesized	4026	102	446	660	1252	2532	5482	10715	12028	14206
defer_der_nonfica_1987	completed	6194	364	364	548	1369	2286	9771	16975	16975	16975
defer_der_nonfica_1987	synthesized	3402	709	709	948	2043	3435	4719	5894	5894	6870
defer_der_nonfica_1988	completed	1761	220	220	220	331	407	1741	6606	10212	10212
defer_der_nonfica_1988	synthesized	1798	509	509	509	875	1235	2582	4203	4203	4371
defer_der_nonfica_1989	completed	1872	41	72	783	1150	1600	1950	3600	3802	8200
defer_der_nonfica_1989	synthesized	1929	82	476	686	1158	1759	2297	3584	4493	5217
defer_der_nonfica_1990	completed	3884	76	293	480	964	1885	3401	6779	7995	12567
defer_der_nonfica_1990	synthesized	2972	40	228	430	916	1752	3066	5593	7638	11695
defer_der_nonfica_1991	completed	2586	25	191	354	836	1779	3179	6309	7500	12254
defer_der_nonfica_1991	synthesized	2523	72	266	436	919	1725	3091	5903	7863	11976
defer_der_nonfica_1992	completed	2472	7	67	224	737	1701	3209	6000	7500	11673
defer_der_nonfica_1992	synthesized	2342	44	172	306	724	1539	2980	5640	7745	11664
defer_der_nonfica_1993	completed	2531	5	60	206	686	1648	3322	6052	7500	12500

Table 61: Percentiles of Synthetic and Completed Variables

Panel 5

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
defer_der_nonfica_1993	synthesized	2319	40	168	302	694	1498	2886	5552	7495	12266
defer_der_nonfica_1994	completed	2448	6	81	199	650	1622	3158	6000	7500	11757
defer_der_nonfica_1994	synthesized	2210	39	153	264	620	1415	2823	5431	7387	11085
defer_der_nonfica_1995	completed	2485	12	81	194	650	1684	3203	6151	7500	12231
defer_der_nonfica_1995	synthesized	2221	41	151	262	597	1418	2818	5465	7431	11431
defer_der_nonfica_1996	completed	2479	7	60	208	650	1800	3340	6031	7500	11114
defer_der_nonfica_1996	synthesized	2330	49	165	286	642	1521	3047	5791	7580	11271
defer_der_nonfica_1997	completed	2671	7	76	225	774	1847	3506	6667	7836	12000
defer_der_nonfica_1997	synthesized	2537	45	181	318	731	1705	3268	6258	8012	11650
defer_der_nonfica_1998	completed	2799	11	97	258	844	1939	3762	7008	8000	12240
defer_der_nonfica_1998	synthesized	2445	43	171	308	709	1640	3147	6108	7841	11298
defer_der_nonfica_1999	completed	2875	11	101	236	817	1971	3874	7125	8035	12240
defer_der_nonfica_1999	synthesized	2507	44	173	307	719	1672	3253	6203	8025	11337
defer_der_nonfica_2000	completed	2927	12	109	267	851	2093	3922	7344	8004	12000
defer_der_nonfica_2000	synthesized	2573	51	193	338	780	1775	3345	6283	8093	11036
defer_der_nonfica_2001	completed	3142	21	120	269	825	2166	4305	7895	8585	12542
defer_der_nonfica_2001	synthesized	2731	50	215	376	821	1816	3606	6892	8520	11118
defer_der_nonfica_2002	completed	3552	17	144	327	968	2400	4928	8500	11000	13995
defer_der_nonfica_2002	synthesized	3158	65	235	398	874	1953	4391	8045	10175	12641
defer_der_nonfica_2003	completed	3777	21	139	300	897	2400	5335	9361	12000	14632
defer_der_nonfica_2003	synthesized	3523	73	264	452	986	2155	4772	8597	11214	14470
SER Earnings Arrays											
earn1937_to_1951	completed	5555	9	58	152	684	2863	8072	14812	19520	30424
earn1937_to_1951	synthesized	6081	3	11	22	150	1662	8508	18553	26658	39760
totearn_ser_1951	completed	1659	9	48	112	443	1463	2821	3600	3600	3600
totearn_ser_1951	synthesized	1528	15	69	134	370	1193	2613	3600	3600	3600
totearn_ser_1952	completed	1754	10	52	123	460	1607	3085	3600	3600	3600
totearn_ser_1952	synthesized	1623	19	81	153	397	1342	2832	3600	3600	3600
totearn_ser_1953	completed	1856	10	55	138	521	1771	3360	3600	3600	3600
totearn_ser_1953	synthesized	1730	21	90	166	439	1490	3137	3600	3600	3600
totearn_ser_1954	completed	1895	10	54	133	537	1851	3478	3600	3600	3600
totearn_ser_1954	synthesized	1764	22	96	178	476	1552	3172	3600	3600	3600
totearn_ser_1955	completed	2060	14	66	152	550	1878	3664	4200	4200	4200
totearn_ser_1955	synthesized	1933	25	105	195	509	1645	3358	4200	4200	4200
totearn_ser_1956	completed	2176	14	74	171	645	2080	3916	4200	4200	4200
totearn_ser_1956	synthesized	2078	27	118	218	595	1883	3667	4200	4200	4200
totearn_ser_1957	completed	2252	16	87	214	774	2190	4065	4200	4200	4200
totearn_ser_1957	synthesized	2159	32	134	251	703	1998	3790	4200	4200	4200
totearn_ser_1958	completed	2282	14	83	203	775	2239	4196	4200	4200	4200
totearn_ser_1958	synthesized	2159	29	125	235	695	2077	3736	4129	4188	4268
totearn_ser_1959	completed	2505	17	92	219	813	2403	4463	4800	4800	4800
totearn_ser_1959	synthesized	2400	30	129	243	714	2217	4181	4800	4800	4800
totearn_ser_1960	completed	2559	18	103	238	857	2473	4629	4800	4800	4800
totearn_ser_1960	synthesized	2438	35	142	261	728	2253	4346	4800	4800	4800
totearn_ser_1961	completed	2602	20	107	247	894	2544	4782	4800	4800	4800
totearn_ser_1961	synthesized	2438	31	131	248	714	2233	4403	4800	4800	4800
totearn_ser_1962	completed	2675	18	107	274	955	2667	4800	4800	4800	4800

Table 61: Percentiles of Synthetic and Completed Variables

Panel 6

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
toteam_ser_1962	synthesized	2532	38	154	284	805	2410	4557	4800	4800	4800
toteam_ser_1963	completed	2718	19	112	270	996	2766	4800	4800	4800	4800
toteam_ser_1963	synthesized	2584	35	148	278	826	2525	4623	4800	4800	4800
toteam_ser_1964	completed	2794	20	117	286	1027	2937	4800	4800	4800	4800
toteam_ser_1964	synthesized	2623	32	138	264	818	2624	4738	4800	4800	4800
toteam_ser_1965	completed	2833	23	128	291	1048	3026	4800	4800	4800	4800
toteam_ser_1965	synthesized	2691	38	159	296	868	2739	4800	4800	4800	4800
toteam_ser_1966	completed	3383	23	132	312	1088	3146	6010	6600	6600	6600
toteam_ser_1966	synthesized	3190	37	162	316	938	2857	5522	6600	6600	6600
toteam_ser_1967	completed	3479	24	137	335	1171	3307	6268	6600	6600	6600
toteam_ser_1967	synthesized	3303	46	188	349	997	2996	5876	6600	6600	6600
toteam_ser_1968	completed	3897	28	156	361	1266	3582	6736	7800	7800	7800
toteam_ser_1968	synthesized	3682	49	202	376	1081	3247	6274	7800	7800	7800
toteam_ser_1969	completed	4057	29	176	398	1350	3818	7248	7800	7800	7800
toteam_ser_1969	synthesized	3880	49	209	402	1173	3525	6817	7800	7800	7800
toteam_ser_1970	completed	4213	28	177	401	1436	4106	7660	7800	7800	7800
toteam_ser_1970	synthesized	4014	53	222	422	1243	3765	7098	7800	7800	7800
toteam_ser_1971	completed	4339	28	178	411	1502	4368	7800	7800	7800	7800
toteam_ser_1971	synthesized	4129	53	221	424	1267	3982	7399	7800	7800	7800
toteam_ser_1972	completed	4794	30	192	444	1595	4618	8683	9000	9000	9000
toteam_ser_1972	synthesized	4553	55	232	448	1348	4244	8055	9000	9000	9000
toteam_ser_1973	completed	5398	38	210	478	1699	4968	9327	10800	10800	10800
toteam_ser_1973	synthesized	5175	60	259	496	1487	4660	8874	10800	10800	10800
toteam_ser_1974	completed	6087	36	226	533	1910	5346	9968	13200	13200	13200
toteam_ser_1974	synthesized	5823	63	289	565	1676	5026	9576	13200	13200	13200
toteam_ser_1975	completed	6458	43	243	551	1965	5683	10626	14100	14100	14100
toteam_ser_1975	synthesized	6194	75	315	601	1748	5333	10190	14100	14100	14100
toteam_ser_1976	completed	7001	45	272	614	2164	6122	11483	15300	15300	15300
toteam_ser_1976	synthesized	6720	82	356	677	1929	5724	11067	15300	15300	15300
toteam_ser_1977	completed	7515	48	287	657	2316	6540	12352	16500	16500	16500
toteam_ser_1977	synthesized	7034	85	366	699	1970	5809	11533	16500	16500	16500
SIPP Arrays											
famwelamt1990	completed	2177	25	113	210	531	1194	2847	5722	7763	11454
famwelamt1990	synthesized	2312	42	172	298	624	1287	2979	5931	8020	12183
famwelamt1991	completed	2664	36	154	287	668	1540	3625	6812	8904	13158
famwelamt1991	synthesized	2238	27	124	228	519	1142	2725	6047	8312	13090
famwelamt1992	completed	2326	21	103	201	515	1274	3156	6093	8074	12301
famwelamt1992	synthesized	2208	28	125	229	514	1139	2769	5887	8075	12760
famwelamt1993	completed	2469	26	116	226	570	1375	3412	6491	8471	12393
famwelamt1993	synthesized	2489	37	154	275	601	1317	3219	6679	9015	13206
famwelamt1994	completed	2264	20	106	207	525	1249	3007	5960	7793	11811
famwelamt1994	synthesized	2096	27	120	217	477	1034	2490	5715	7932	12906
famwelamt1995	completed	1920	21	98	193	487	1082	2276	5090	7183	10815
famwelamt1995	synthesized	2102	28	119	214	478	1047	2406	5881	8166	13120
famwelamt1996	completed	2888	33	127	244	621	1525	4015	7476	9846	15097
famwelamt1996	synthesized	3026	46	185	319	683	1489	3912	7973	10789	17421
famwelamt1997	completed	2594	28	121	224	561	1339	3332	7025	9424	13458

Table 61: Percentiles of Synthetic and Completed Variables

Panel 7

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
famwelamt1997	synthesized	2999	40	165	286	616	1377	3681	8443	11065	19387
famwelamt1998	completed	2243	23	103	193	489	1169	2797	6075	8256	12887
famwelamt1998	synthesized	2546	36	139	239	510	1138	2927	7273	10418	16405
famwelamt1999	completed	1940	27	116	203	474	1033	2294	4948	7294	12055
famwelamt1999	synthesized	2344	39	151	256	524	1116	2724	6466	9024	14643
fpov1990	completed	129678	73957	79746	84031	99608	119033	157243	187777	208340	272232
fpov1990	synthesized	131862	74677	80581	86047	101730	121391	158244	188948	208597	267694
fpov1991	completed	135912	76675	82173	86352	103824	125006	165633	197304	218963	279869
fpov1991	synthesized	140209	77183	83287	89093	106564	128769	168589	201333	223398	283160
fpov1992	completed	139175	79284	84093	88308	106343	127602	170288	202009	224766	290261
fpov1992	synthesized	143818	79805	85760	91731	109769	132306	173147	206196	227965	291539
fpov1993	completed	143604	81782	87016	90732	109186	131548	175718	209927	232273	300478
fpov1993	synthesized	146804	82261	87906	93057	111411	134799	177797	212475	234327	296431
fpov1994	completed	146660	81477	87869	92505	110820	134750	180044	213536	235850	299578
fpov1994	synthesized	148762	82375	88859	94387	112734	137094	181409	215045	235848	295635
fpov1995	completed	149438	79258	88736	95207	111751	135717	183384	218590	241636	314551
fpov1995	synthesized	152346	80973	90013	96712	113828	139191	185755	221023	244702	315943
fpov1996	completed	155213	87867	92638	97883	114286	139828	191028	230074	255309	333004
fpov1996	synthesized	157046	88733	94111	99447	116965	142602	192227	230007	253485	323077
fpov1997	completed	154949	88870	93505	99683	116322	139515	194903	230144	255571	321650
fpov1997	synthesized	160714	90227	95335	100821	118977	146171	197707	236812	261760	329778
fpov1998	completed	158996	89096	94824	101382	118264	145376	198289	234755	259364	323841
fpov1998	synthesized	162224	90015	96439	102150	120532	147505	199928	239316	264133	334169
fpov1999	completed	162718	92542	97229	103731	120771	148233	202576	240540	265306	334363
fpov1999	synthesized	168355	92148	98489	104493	123625	152739	205995	248062	276256	358243
ftotinc1990	completed	36965	1044	5735	9236	17692	31328	49528	71261	88542	128397
ftotinc1990	synthesized	35271	658	5548	8978	17156	30137	47206	67433	83269	121216
ftotinc1991	completed	38713	1764	6461	9988	18967	33325	52288	74401	90330	123523
ftotinc1991	synthesized	35629	-68	4743	8000	15997	29520	48308	71115	88221	125445
ftotinc1992	completed	38913	1301	6146	9669	18636	33094	52881	75954	92174	125975
ftotinc1992	synthesized	36815	22	5122	8532	16863	30669	49966	73493	90209	126303
ftotinc1993	completed	40269	2080	6921	10567	19555	34437	54375	77512	94570	129698
ftotinc1993	synthesized	37969	85	5283	8810	17345	31781	51513	75370	92961	129897
ftotinc1994	completed	41783	1560	6700	10508	19767	35334	56565	82064	100287	138129
ftotinc1994	synthesized	38788	-194	5142	8751	17337	32109	52991	78275	96870	136357
ftotinc1995	completed	43020	1263	6615	10793	20432	37055	58670	83405	100230	142381
ftotinc1995	synthesized	40829	-997	4067	7914	17084	33432	56576	83688	103719	147841
ftotinc1996	completed	46703	1094	6511	10421	20293	36916	60535	89881	112724	208588
ftotinc1996	synthesized	43897	-87	5679	9580	18934	34908	57415	85490	107185	188185
ftotinc1997	completed	46358	1219	6322	9954	19464	36169	60471	90228	114300	213361
ftotinc1997	synthesized	45180	112	5671	9494	18858	35078	58859	88490	112590	208202
ftotinc1998	completed	48992	1235	6623	10482	20547	38283	64043	95603	120897	231081
ftotinc1998	synthesized	48276	84	5885	9897	19792	37109	62751	95196	121927	239190
ftotinc1999	completed	52561	1326	6734	10858	21539	40509	67285	101817	133239	274523
ftotinc1999	synthesized	51342	-389	5574	9789	20193	38718	65731	101291	133760	282312
helamt1990	completed	2213	38	154	285	592	1148	2399	5048	7983	17238
helamt1990	synthesized	2051	32	144	256	538	1030	2094	4526	7480	17343

Table 61: Percentiles of Synthetic and Completed Variables

Panel 8

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
helamt1991	completed	2867	51	221	396	814	1527	3211	6527	9840	20633
helamt1991	synthesized	2886	57	228	394	789	1487	3048	6611	10327	22101
helamt1992	completed	3189	55	211	388	834	1688	3507	7507	11287	21015
helamt1992	synthesized	2974	51	219	392	795	1530	3109	6591	10669	22558
helamt1993	completed	2917	47	194	360	794	1598	3327	6650	9776	20823
helamt1993	synthesized	3021	55	226	395	804	1554	3201	6768	10773	22825
helamt1994	completed	3269	56	227	408	897	1774	3791	7514	11174	22554
helamt1994	synthesized	3315	56	235	416	869	1697	3631	7606	12001	25077
helamt1995	completed	3965	116	427	706	1346	2518	4694	8655	13187	21299
helamt1995	synthesized	3594	84	299	514	1066	2085	4035	7768	12943	23200
helamt1996	completed	3287	54	243	434	903	1797	3577	6740	10479	27748
helamt1996	synthesized	3844	83	308	517	985	1842	3735	7848	13695	37446
helamt1997	completed	4102	84	296	512	1048	2111	4563	8925	13660	24256
helamt1997	synthesized	6215	130	436	702	1295	2510	5459	11471	18987	33202
helamt1998	completed	4784	113	360	637	1189	2347	4757	9745	14310	26837
helamt1998	synthesized	6500	113	420	673	1266	2403	4768	10832	18666	89301
helamt1999	completed	4945	119	446	727	1329	2587	5098	9523	14210	26510
helamt1999	synthesized	7380	120	467	763	1453	2743	5604	11665	20054	85726
totearn1990	completed	17643	107	697	1587	5355	13886	24881	37967	48411	77399
totearn1990	synthesized	16465	85	562	1140	3881	12559	23673	36576	46880	75989
totearn1991	completed	18506	148	832	1768	5549	14336	25948	40109	51367	83616
totearn1991	synthesized	17662	95	579	1242	4300	13236	25163	39441	50893	84192
totearn1992	completed	19353	158	918	1985	6040	14917	26982	41831	54170	87016
totearn1992	synthesized	17881	85	614	1378	4579	13343	25324	39815	51771	85377
totearn1993	completed	19519	151	915	1966	6086	15059	27189	42143	54978	87849
totearn1993	synthesized	19016	99	631	1397	4812	14218	26950	42415	55377	90114
totearn1994	completed	19871	197	1084	2231	6348	15214	27534	42916	55686	90189
totearn1994	synthesized	19076	112	702	1553	5061	14260	26869	42468	55240	90622
totearn1995	completed	19857	285	1525	2943	7014	15132	27000	42197	54303	91142
totearn1995	synthesized	19385	346	1568	2786	6364	14103	26438	42250	55008	92325
totearn1996	completed	24323	176	1197	2574	7483	17418	31521	49456	66129	138332
totearn1996	synthesized	24352	231	1198	2457	7136	17165	31517	49558	66508	145724
totearn1997	completed	24571	233	1313	2811	7870	17855	31911	49723	65322	130138
totearn1997	synthesized	25592	219	1172	2465	7498	18027	32874	51952	69279	151486
totearn1998	completed	25351	267	1544	3196	8459	18560	32758	51483	67395	133194
totearn1998	synthesized	27054	233	1285	2732	8136	18971	34317	55186	73731	169917
totearn1999	completed	27834	354	2006	3947	9661	20043	35050	55260	73703	171998
totearn1999	synthesized	29130	159	1205	2811	8574	19891	36079	58452	80580	213952
tohoursannual1990	completed	1671	44	181	346	972	1945	2197	2595	2915	3537
tohoursannual1990	synthesized	1521	19	103	220	729	1801	2141	2435	2691	3257
tohoursannual1991	completed	1603	47	178	327	852	1871	2145	2527	2843	3608
tohoursannual1991	synthesized	1511	19	115	235	724	1762	2129	2435	2717	3317
tohoursannual1992	completed	1683	55	216	419	1072	1954	2169	2517	2812	3450
tohoursannual1992	synthesized	1561	40	169	308	822	1802	2140	2470	2762	3403
tohoursannual1993	completed	1706	53	205	386	1010	1970	2208	2673	3003	3805
tohoursannual1993	synthesized	1584	43	167	303	815	1836	2166	2542	2842	3478
tohoursannual1994	completed	1651	64	226	405	976	1923	2153	2558	2855	3478

Table 61: Percentiles of Synthetic and Completed Variables

Panel 9

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
tohoursannual1994	synthesized	1557	68	219	368	838	1762	2128	2460	2749	3332
tohoursannual1995	completed	1620	111	320	506	972	1836	2135	2452	2726	3243
tohoursannual1995	synthesized	1550	147	366	545	962	1677	2089	2318	2542	3044
tohoursannual1996	completed	1743	61	235	439	1117	1965	2242	2702	3015	3648
tohoursannual1996	synthesized	1711	67	239	431	1078	1929	2226	2645	2943	3546
tohoursannual1997	completed	1730	57	234	445	1116	1985	2200	2631	2946	3644
tohoursannual1997	synthesized	1700	48	194	370	1027	1951	2218	2645	2957	3665
tohoursannual1998	completed	1726	61	238	464	1117	1986	2182	2608	2923	3587
tohoursannual1998	synthesized	1711	56	216	413	1069	1954	2213	2623	2932	3632
tohoursannual1999	completed	1747	69	277	526	1182	1995	2182	2597	2888	3553
tohoursannual1999	synthesized	1690	45	201	393	999	1924	2202	2655	3002	3778
totinc1990	completed	16615	-829	152	1024	4628	12676	23770	36793	47160	74068
totinc1990	synthesized	16282	-764	99	827	3872	11736	23480	37345	48158	76832
totinc1991	completed	17284	-675	200	1021	4762	12941	24587	38580	49573	80636
totinc1991	synthesized	16815	-569	268	896	3941	11756	24041	38863	50491	83379
totinc1992	completed	17762	-720	192	1100	5050	13087	25042	39776	51663	84390
totinc1992	synthesized	17468	-511	324	1016	4408	12132	24663	40284	52919	87698
totinc1993	completed	18300	-667	351	1490	5493	13637	25608	40420	52413	85707
totinc1993	synthesized	18065	-494	457	1330	4973	12830	25340	40916	53532	88560
totinc1994	completed	18667	-683	734	2040	5966	13942	25866	41116	53059	86348
totinc1994	synthesized	18714	-487	769	1825	5358	13212	25979	42485	55541	91004
totinc1995	completed	18995	-758	1081	2627	6400	14060	26131	41229	52884	87577
totinc1995	synthesized	18996	-636	1076	2364	5699	13036	25962	42851	56489	94769
totinc1996	completed	21841	-764	893	2525	6892	15523	28768	45732	59370	112062
totinc1996	synthesized	22011	-322	996	2411	6579	15158	28879	46644	61269	116830
totinc1997	completed	23103	-290	1089	2873	7444	16552	30285	47749	62296	117509
totinc1997	synthesized	22751	-258	1091	2496	6809	15675	29693	47855	63014	122999
totinc1998	completed	23804	-399	1297	3233	7895	17292	31288	49679	64701	115910
totinc1998	synthesized	23799	-282	1359	2901	7268	16360	30713	50201	66632	130520
totinc1999	completed	26042	-373	1678	3669	8520	18087	32969	52738	70720	163434
totinc1999	synthesized	25988	-207	1663	3363	7801	17205	32788	53786	72632	171044
wkspt1990	completed	15.8	0.3	1.1	2.2	5.2	12.3	22.5	36.7	44.5	52
wkspt1990	synthesized	15.7	0.2	1.2	2.2	5.1	12	22.4	36.7	44.2	51.1
wkspt1991	completed	15.8	0.3	1.1	2.1	5	12.1	22.5	37.1	45.7	52
wkspt1991	synthesized	15.7	0.3	1.2	2.2	5.1	11.9	22.2	37	45.2	52
wkspt1992	completed	16.9	0.3	1.1	2.2	5.6	13.4	24.5	38.2	45.8	53
wkspt1992	synthesized	16.3	0.2	1.2	2.2	5.4	12.8	23.6	37.2	44.3	52
wkspt1993	completed	16.4	0.3	1.1	2.2	5.3	12.8	23.6	37.8	46.7	52
wkspt1993	synthesized	16.2	0.3	1.2	2.3	5.3	12.4	23.1	38	46.5	52
wkspt1994	completed	15.7	0.3	1.2	2.1	5.1	12.4	21.4	36.5	45.7	52
wkspt1994	synthesized	15.9	0.3	1.4	2.5	5.5	12.7	21.6	36.2	44.7	52
wkspt1995	completed	16.5	0.6	2.2	3.8	8.2	13.9	23.5	32.4	40	49.4
wkspt1995	synthesized	18.1	0.7	2.3	3.9	8.5	15.7	25.1	36.6	42.2	49.5
wkspt1996	completed	22.3	0.7	2.8	4.9	11.2	19.1	33.2	44.6	48.8	52
wkspt1996	synthesized	22	0.5	2.3	4.2	10.4	18.9	33.4	44.8	48.6	51.9
wkspt1997	completed	19.8	0.4	1.7	3.3	8.1	16.5	29.7	43.2	49.4	53
wkspt1997	synthesized	20.6	0.4	1.9	3.5	8.3	17.2	31.4	44.6	49.8	52

Table 61: Percentiles of Synthetic and Completed Variables

Panel 10

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
wkspt1998	completed	20	0.6	2.2	3.9	8.3	16.5	29.7	43	49.4	52
wkspt1998	synthesized	21.2	0.6	2.5	4.3	9.2	17.9	32	45	49.9	52
wkspt1999	completed	20	0.5	2.2	3.8	8.6	16.8	29.7	42.3	48.5	52
wkspt1999	synthesized	21.9	0.6	2.7	4.7	10	18.8	32.9	45	49.3	52
wkswp1990	completed	41	3.2	10	17.3	35.3	47.5	51.2	51.9	52	52
wkswp1990	synthesized	40.7	2	7.5	13.8	34.7	48.5	51.2	51.9	52	52
wkswp1991	completed	41.5	3.6	10.4	17.2	35.7	48.8	51.6	52	52	52
wkswp1991	synthesized	41.2	2.6	9.1	15.7	34.5	49	51.5	52	52	52
wkswp1992	completed	41.9	3.8	11.5	19.4	36.9	47.6	51.5	53	53	53
wkswp1992	synthesized	41.5	2.5	9.5	17.1	36.7	48.2	51.3	52	52	52
wkswp1993	completed	42.4	3.8	11.1	18.5	38.1	49.3	51.7	52	52	52
wkswp1993	synthesized	42.1	2.6	9.6	16.9	37.1	49.7	51.6	52	52	52
wkswp1994	completed	42	4.5	12.8	19.3	36.3	48.5	51.6	52	52	52
wkswp1994	synthesized	42.2	3.9	12.1	19.1	36.9	49.1	51.4	51.9	52	52
wkswp1995	completed	41.2	6.8	15.1	20.2	35.4	46.5	50.5	51.6	51.8	52
wkswp1995	synthesized	41.3	4.5	12.4	19.4	35.4	47.5	50.7	51.6	51.8	52
wkswp1996	completed	43.2	4	13.5	22.4	39.7	49.3	51.5	52	52	52
wkswp1996	synthesized	44	3.7	13.6	24	42	49.9	51.5	52	52	52
wkswp1997	completed	43.3	3.6	11.5	20	39	50.9	51.9	52	53	53
wkswp1997	synthesized	43.8	2.8	11.1	20.7	41.6	50.8	51.9	52	52	52
wkswp1998	completed	44	4.3	14.9	24.9	40.9	50.1	51.9	52	52	52
wkswp1998	synthesized	44.7	4.7	15.5	26.4	43.1	50.5	51.8	52	52	52
wkswp1999	completed	43.9	5.8	16	26.1	40.4	49.5	51.8	52	52	52
wkswp1999	synthesized	44.8	4.7	16.1	27.2	43	50.5	51.8	52	52	52

Table 61: Percentiles of Synthetic and Completed Variables

Panel 11

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
Cardinal Categorical Variables											
time_arrive_usa	completed	5.56	1	1	2	4	6	8	8	8	8
time_arrive_usa	synthesized	5.44	1	1	1.19	4	6	8	8	8	8
totfam_kids	completed	0.9	0	0	0	0	0	2	3	3	5
totfam_kids	synthesized	0.9	0	0	0	0	0	2	3	3	5

Table 62: Selected Variables – Weighted and unweighted counts and percentages (averaged across completed and synthetic implicates)

Variable	WEIGHTED				UNWEIGHTED			
	Count		Percentage		Count		Percentage	
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed
Male								
0	106,949,968	106,470,823	52.41	52.18	138,357	138,357	52.45	52.45
1	97,094,759	97,573,904	47.59	47.82	125,436	125,436	47.55	47.55
Black								
0	181,725,962	180,358,529	89.06	88.39	232,401	233,326	88.10	88.45
1	22,318,765	23,686,198	10.94	11.61	31,392	30,467	11.90	11.55
Hispanic								
0	184,181,614	181,560,851	90.27	88.98	238,277	238,558	90.33	90.43
1	19,863,113	22,483,876	9.73	11.02	25,516	25,235	9.67	9.57
Maritalstat								
1	103,680,995	102,213,429	50.81	50.09	141,292	141,292	53.56	53.56
2	9,673,233	9,958,884	4.74	4.88	17,483	17,483	6.63	6.63
3	23,885,992	24,400,916	11.71	11.96	31,103	31,103	11.79	11.79
4	66,804,507	67,471,498	32.74	33.07	73,915	73,915	28.02	28.02
Tob_initial								
	144,158,303	142,951,882	70.65	70.06	173,251	173,251	65.68	65.68
1	28,120,067	28,896,858	13.78	14.16	45,543	45,543	17.26	17.26
2	9,771,612	9,934,114	4.79	4.87	15,401	15,401	5.84	5.84
3	2,562,145	2,603,705	1.26	1.28	4,028	4,028	1.53	1.53
5	3,881,472	4,000,409	1.90	1.96	6,920	6,920	2.62	2.62
100	15,551,128	15,657,759	7.62	7.67	18,652	18,652	7.07	7.07
Tob_2000								
	150,570,926	149,502,998	73.79	73.27	196,028	196,028	74.31	74.31
1	26,411,823	27,162,166	12.94	13.31	34,306	34,306	13.00	13.00
2	5,873,327	5,926,142	2.88	2.90	7,571	7,571	2.87	2.87
3	1,860,289	1,880,083	0.91	0.92	2,433	2,433	0.92	0.92
5	4,311,070	4,460,705	2.11	2.19	5,781	5,781	2.19	2.19
100	15,017,292	15,112,633	7.36	7.41	17,675	17,675	6.70	6.70
Own_home								
0	69,152,585	69,785,113	33.89	34.20	87,230	86,839	33.07	32.92
1	134,892,142	134,259,614	66.11	65.80	176,563	176,955	66.93	67.08
Foreign_born								
0	180,969,389	179,239,395	88.69	87.84	234,496	235,262	88.89	89.18
1	23,075,338	24,805,332	11.31	12.16	29,297	28,531	11.11	10.82

Table 62: Selected Variables – Weighted and unweighted counts and percentages (averaged across completed and synthetic implicates)

Variable	WEIGHTED				UNWEIGHTED				
	Count		Percentage		Count		Percentage		
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	
Educ_5cat									
1	43,271,926	42,898,854	21.21	21.02	56,851	55,903	21.55	21.19	
2	69,217,821	68,520,001	33.92	33.58	89,529	90,001	33.94	34.12	
3	51,100,854	52,842,654	25.04	25.90	64,585	66,115	24.48	25.06	
4	22,930,251	23,458,739	11.24	11.50	29,804	29,230	11.30	11.08	
5	17,523,875	16,324,479	8.59	8.00	23,024	22,544	8.73	8.55	
Age_cat12									
<=21	10,993,657	12,355,657	5.39	6.06	4,452	4,480	1.69	1.70	
22-24	11,083,762	11,637,091	5.43	5.70	8,296	8,484	3.14	3.22	
25-29	19,330,649	18,791,213	9.47	9.21	24,922	25,344	9.45	9.61	
30-34	20,827,765	19,839,766	10.21	9.72	28,311	27,657	10.73	10.48	
35-39	22,623,857	22,137,354	11.09	10.85	28,824	28,715	10.93	10.89	
40-44	22,702,993	22,384,522	11.13	10.97	28,493	28,579	10.80	10.83	
45-49	20,541,360	20,091,637	10.07	9.85	26,195	26,150	9.93	9.91	
50-54	18,114,622	17,779,895	8.88	8.71	23,761	23,797	9.01	9.02	
55-59	13,593,649	13,497,244	6.66	6.61	17,822	17,895	6.76	6.78	
60-64	10,666,518	10,893,885	5.23	5.34	14,636	14,469	5.55	5.48	
65-69	9,336,754	9,547,527	4.58	4.68	12,829	12,849	4.86	4.87	
>=70	24,229,138	25,088,937	11.87	12.30	45,255	45,375	17.16	17.20	
Agecat_initial_entitle (Tob_initial=1,2,3,5)									
<62	11,699,746	11,688,294	26.39	25.73	18,555	18,184	25.81	25.29	
>=62 and <63	7,273,818	18,508,709	16.41	40.74	11,673	29,067	16.24	40.43	
>=63 and <64	11,912,607	3,419,745	26.87	7.53	19,452	5,391	27.06	7.50	
>=64 and <65	7,337,083	2,757,925	16.55	6.07	11,964	4,961	16.64	6.90	
>=65 and <66	3,351,021	7,440,877	7.56	16.38	5,536	11,672	7.70	16.24	
>=66 and <67	1,514,780	482,726	3.42	1.06	2,587	798	3.60	1.11	
>=67	1,208,960	1,077,981	2.73	2.37	2,051	1,726	2.85	2.40	
>=80	37,283	58,829	0.08	0.13	73	90	0.10	0.13	
Agecat_retire (Tob_initial=1 only)									
<62	56,373	45,537	0.20	0.16	83	73	0.18	0.16	
>=62 and <63	5,843,165	15,661,703	20.78	54.20	9,187	24,252	20.17	53.25	
>=63 and <64	10,449,792	2,770,806	37.16	9.59	16,945	4,374	37.21	9.60	
>=64 and <65	6,457,351	2,351,612	22.96	8.14	10,486	4,237	23.02	9.30	
>=65 and <66	2,941,391	6,703,497	10.46	23.20	4,850	10,471	10.65	22.99	
>=66 and <67	1,323,219	406,445	4.71	1.41	2,259	663	4.96	1.46	
>=67	1,020,418	910,199	3.63	3.15	1,677	1,403	3.68	3.08	
>=80	28,357	47,059	0.10	0.16	56	71	0.12	0.16	

Table 63: Agreement Probabilities for Individuals with Spouses

Field	Comparison Type	Pr(agree match): m	Pr(agree non-match): u	Agree weight: $\ln(m/u)$	Disagree weight: $\ln(1-m)/(1-u)$
Hispanic	c	0.954479	0.835287	0.133390	-1.286023
Educ_5cat	c	0.330004	0.241200	0.313478	-0.124467
Disab_in_scope	c	0.949006	0.777256	0.199645	-1.474307
Disab	c	0.843075	0.810676	0.039187	-0.187691
Disab_nowork	c	0.637131	0.541970	0.161765	-0.232893
Totfam_kids_wave2	c	0.469601	0.329187	0.355257	-0.234861
Ind_4cat	c	0.361122	0.309276	0.154980	-0.078026
Foreign_born	c	0.844434	0.788724	0.068250	-0.306097
Time_arrive_usa	c	0.236797	0.162303	0.377738	-0.093133
Ind_exist	c	0.762450	0.568762	0.293074	-0.596280
Occ_exist	c	0.775007	0.572171	0.303434	-0.642654
Occ_4cat	c	0.446905	0.343057	0.264449	-0.172067
Mh_category	c	0.591162	0.574111	0.029268	-0.040861
Flag_mar4t	c	0.987294	0.987260	0.000035	-0.002695
Own_home	c	0.719070	0.668007	0.073660	-0.167008
Pension_in_scope_age	c	0.976252	0.949419	0.027870	-0.756061
Pension in scope empl	c	0.702327	0.557740	0.230506	-0.395902

Table 64: Agreement Probabilities for Single Individuals

Field	Comparison Type	Pr(agree match): m	Pr(agree non-match): u	Agree weight: $\ln(m/u)$	Disagree weight: $\ln(1-m)/(1-u)$
Hispanic	c	0.888222	0.817697	0.082729	-0.489153
Educ_5cat	c	0.360123	0.252198	0.356231	-0.155862
Disab_in_scope	c	0.923310	0.744927	0.214679	-1.201784
Disab	c	0.824805	0.113998	1.978968	-1.620817
Disab_nowork	c	0.679595	0.222995	1.114350	-0.885862
Totfam_kids_wave2	c	0.568113	0.130233	1.472992	-0.700061
Ind_4cat	c	0.356281	0.305685	0.153165	-0.075664
Foreign_born	c	0.852712	0.094033	2.204775	-1.816610
Time_arrive_usa	c	0.289757	0.091983	1.147440	-0.245656
Ind_exist	c	0.784428	0.603121	0.262838	-0.610339
Occ_exist	c	0.784490	0.602726	0.263572	-0.611621
Occ_4cat	c	0.465897	0.388607	0.181394	-0.135150
Mh_category	c	0.763459	0.067933	2.419334	-1.371281
Flag_mar4t	c	0.990087	0.004855	5.317686	-4.609064
Own_home	c	0.547307	0.242271	0.814954	-0.515111
Pension_in_scope_age	c	0.887510	0.585350	0.416210	-1.304568
Pension in scope empl	c	0.693329	0.211577	1.186915	-0.944258

Table 65: Match Rates for Married Individuals, Split into Data Blocks

<i>Segment</i>	<i>Match Status</i>	<i>COUNT</i>	<i>PERCENT</i>
1	FALSE	29939	99.31
	TRUE	209	0.69
2	FALSE	19660	99.57
	TRUE	84	0.43
3	FALSE	19517	99.62
	TRUE	74	0.38
4	FALSE	20202	99.71
	TRUE	58	0.29
5	FALSE	20017	99.71
	TRUE	58	0.29
6	FALSE	19811	99.62
	TRUE	76	0.38
7	FALSE	19658	99.65
	TRUE	69	0.35
8	FALSE	19564	99.7
	TRUE	58	0.3
9	FALSE	18305	99.63
	TRUE	68	0.37
10	FALSE	19724	99.73
	TRUE	54	0.27

Table 66: Match Rates for Single Individuals, Split into Data Blocks

<i>Segment</i>	<i>matchstatus</i>	<i>COUNT</i>	<i>PERCENT</i>
1	FALSE	21717	99.2
	TRUE	175	0.8
2	FALSE	18005	98.82
	TRUE	215	1.18
3	FALSE	18028	99.1
	TRUE	164	0.9
4	FALSE	18936	99.28
	TRUE	138	0.72
5	FALSE	19102	99.29
	TRUE	136	0.71
6	FALSE	18503	99.18
	TRUE	153	0.82
7	FALSE	18682	99.22
	TRUE	146	0.78
8	FALSE	18798	99.34
	TRUE	124	0.66
9	FALSE	19034	99.3
	TRUE	134	0.7
10	FALSE	19014	99.31
	TRUE	132	0.69
11	FALSE	17411	98.83
	TRUE	207	1.17
12	FALSE	19018	99.25
	TRUE	144	0.75
13	FALSE	29939	99.31
	TRUE	209	0.69

Table 67: Mahalanobis Distance Matching Results

Male	Marital Status	N Synth	N N GS	Match Rate 1 Maha1	Match Rate 2 Maha1	Ratio 2 to 1	Match Rate 3 Maha1	Ratio 3 to 2	Ratio 3, 2 to 1
1	1	70,814	70,814	1.11	0.50	0.45	0.44	0.88	0.84
0	1	70,478	70,478	1.03	0.55	0.53	0.44	0.81	0.96
1	4	39,434	39,434	0.97	0.52	0.54	0.39	0.74	0.93
0	4	34,481	34,481	1.18	0.73	0.62	0.55	0.74	1.09
0	3	18,733	18,733	1.05	0.54	0.51	0.33	0.61	0.83
0	2	14,668	14,668	1.04	0.67	0.64	0.50	0.74	1.12
1	3	12,370	12,370	1.04	0.46	0.44	0.38	0.82	0.81
1	2	2,815	2,815	2.91	1.53	0.52	0.78	0.51	0.79
Totals		263,793	263,793	1.09	0.57	0.52	0.44	0.79	0.93
Male	Marital Status	N Synth	N N GS	Match Rate 1 Maha2	Match Rate 2 Maha2	Ratio 2 to 1	Match Rate 3 Maha2	Ratio 3 to 2	Ratio 3, 2 to 1
1	1	70,814	70,814	0.80	0.39	0.48	0.31	0.81	0.87
0	1	70,478	70,478	0.67	0.38	0.57	0.32	0.83	1.05
1	4	39,434	39,434	0.68	0.39	0.58	0.28	0.71	0.99
0	4	34,481	34,481	0.80	0.50	0.63	0.42	0.84	1.15
0	3	18,733	18,733	0.64	0.40	0.62	0.34	0.85	1.15
0	2	14,668	14,668	0.78	0.41	0.53	0.38	0.93	1.02
1	3	12,370	12,370	0.74	0.30	0.41	0.35	1.16	0.88
1	2	2,815	2,815	2.20	0.99	0.45	0.75	0.75	0.79
Totals		263,793	263,793	0.75	0.41	0.55	0.34	0.83	1.00

Table 68: Euclidean Distance Matching Results

Male	Marital Status	N Synth	N N GS	Match Rate 1 EUCL1	Match Rate 2 EUCL1	Ratio 2 to 1	Match Rate 3 EUCL1	Ratio 3 to 2	Ratio 3, 2 to 1
1	1	70,814	70,814	0.60	0.40	0.66	0.31	0.77	1.17
0	1	70,478	70,478	0.58	0.39	0.67	0.27	0.71	1.15
1	4	39,434	39,434	0.49	0.28	0.58	0.21	0.75	1.01
0	4	34,481	34,481	0.53	0.32	0.61	0.30	0.93	1.18
0	3	18,733	18,733	0.90	0.57	0.63	0.36	0.63	1.03
0	2	14,668	14,668	0.47	0.42	0.90	0.22	0.53	1.38
1	3	12,370	12,370	0.74	0.45	0.61	0.40	0.88	1.14
1	2	2,815	2,815	0.82	0.50	0.61	0.36	0.71	1.04
Totals		263,793	263,793	0.59	0.38	0.65	0.29	0.75	1.14
Male	Marital Status	N Synth	N N GS	Match Rate 1 EUCL2	Match Rate 2 EUCL2	Ratio 2 to 1	Match Rate 3 EUCL2	Ratio 3 to 2	Ratio 3, 2 to 1
1	1	70,814	70,814	1.26	0.74	0.58	0.55	0.75	1.02
0	1	70,478	70,478	1.43	0.81	0.57	0.66	0.81	1.03
1	4	39,434	39,434	0.94	0.59	0.62	0.51	0.87	1.16
0	4	34,481	34,481	1.16	0.67	0.58	0.51	0.76	1.02
0	3	18,733	18,733	0.91	0.56	0.61	0.42	0.76	1.07
0	2	14,668	14,668	1.03	0.53	0.52	0.52	0.99	1.03
1	3	12,370	12,370	0.91	0.53	0.58	0.44	0.85	1.06
1	2	2,815	2,815	2.31	1.17	0.51	1.03	0.88	0.95
Totals		263,793	263,793	1.20	0.70	0.58	0.56	0.81	1.05

This version: \$Id: SSAfinal.tex 188 2006-11-06 01:10:54Z jma7 \$